



11

TRANSACTIONS OF THE ELEVENTH ARMY

CONFERENCE ON APPLIED MATHEMATICS

AND COMPUTING



DT  
ELE  
JUN 06 1994  
S F U

Approved for public release; distribution unlimited.  
The findings in this report are not to be construed as  
an official Department of the Army position, unless  
so designated by other authorized documents.

This Document Contains  
Missing Page/s That Are  
Unavailable In The  
Original Document

Sponsored by

94-16585



The Mathematical and Computer Sciences Division of the

ARMY RESEARCH OFFICE

94 6 3 030

**U.S. ARMY RESEARCH OFFICE**

**Report No. 94-1**

**March 1994**

**TRANSACTIONS OF THE ELEVENTH ARMY CONFERENCE  
ON APPLIED MATHEMATICS AND COMPUTING**

**Sponsored by the Army Mathematics Coordinating Group**

**HOST**

**Carnegie Mellon University  
Pittsburgh, Pennsylvania  
8-10 June 1993**

**Approved for public release; distributions unlimited.  
The findings in this report are not to be constructed as  
an official Department of the Army position, unless so  
designated by other authorized documents.**

**U.S. Army Research Office  
P.O. Box 12211  
Research Triangle Park, NC 27709-2211**

Accession For	
NTIS	CRA&I <input checked="" type="checkbox"/>
DTIC	TAB <input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

**DTIC QUALITY INSPECTED 2**

## FOREWORD

The host of the Eleventh Conference on Applied Mathematics and Computing was the Center for Nonlinear Analysis at the Carnegie-Mellon University, Pittsburgh, P.A. It was held on June 8-10, 1993. Professor Morton E. Gurtin, the Principal Investigator of this Center, served as Chairperson on local arrangements. He was assisted in this task by Ms. Francine Johnson. We would like to take this opportunity to thank these individuals for all the time and work preparing for and conducting this extremely well managed scientific meeting.

The 1993 conference was attended by more than 70 scientists and engineers representing various Army agencies and academe. The meeting featured five plenary talks and four special sessions on topics of current interest, such as Mathematics of Materials, topics in stochastic Analysis, Computational issues in Geosciences, and Virtual Factory. In addition there were 37 contributed papers presented. The names of the invited speakers and the titles of their addresses are listed below.

### SPEAKER AND AFFILIATION

### TITLE OF ADDRESS

Professor John N. Tistsiklis  
Massachusetts Institute of  
Technology

Complexity Theoretic Aspects  
of Problems in Control  
Theory

Professor Roy A. Nicolaides  
Carnegie Mellon University

Numerical Solutions in  
Microstructure and the  
Calculus of Variations

Professor P. S. Krishnaprasad  
University of Maryland

Rational Wavelets in  
Controls

Professor David Yuen  
University of Minnesota

The Role of Hard Thermal  
Convection in Geosciences

Professor Richard Durrett  
Cornell University

Stochastic Spatial Models of  
Epidemics and Excitable Media

Many of the papers given at this conference provided the attendees a chance to see scientific techniques developments taking place in the Army laboratories. Through these meetings techniques developed at one installation are brought to the attention of scientists at other places, thereby reducing duplication of effort. Another important phase of these meetings is presenting the members of the audience an opportunity to hear nationally known scientists discuss recent developments in their own field.

## TABLE OF CONTENTS\*

<u>Title</u>	<u>Page</u>
Foreword .....	iii
Table of Contents .....	v
Agenda .....	ix
Complexity Theoretic Aspects of Problems in Control Theory John N. Tsitsiklis .....	1
Incoherent Solid-Solid Phase Transitions P. Cermelli, and Morton E. Gurtin .....	7
Formation and Development of Shear Bands in Granular Material Michael Shearer, F. Xabier Garaizar, D. G. Shaeffer, and J. Trangenstein .....	15
Toward a Euclidean Algorithm for Composition of Rational Functions Moss Sweedler .....	29
Decoding Hyperbolic Cascaded Reed-Solomon Codes Keith Saints .....	33
On the Computational Content of Classical Sequent Proofs Judith Underwood .....	41
Computing the Newtonian Graph Dexter Kozen and Kjartan Stefansson .....	55
Implementing Mixed Chaining in a Classification Type Expert System Andrew W. Harrell .....	71

\*This Table of Contents lists only the papers that are published in this Technical Manual. For a list of all the papers presented at the tenth Army Conference on Applied Mathematics and Computing, see the Agenda.



<b><u>Title</u></b>	<b><u>Page</u></b>
<b>Theoretical Algorithms for Solving the Army Stationing Problem</b> Janet Hurst Spoonamore .....	109
<b>Severely Constrained Allocation of a Bounded Number of Transceivers</b> T. Cronin .....	115
<b>CEMPLOT: A Visualization Tool for CEM VII</b> K. Shivaswamy, P. Burns, and D. Alciatore .....	129
<b>Theory of a Bifurcating Model Neuron: A Nonlinear Dynamic Systems Approach</b> N. H. Farhat, M. Eldefrawy and S-Y Lin .....	145
<b>Shock Induced Surface Instabilities and Nonlinear Wave Interactions</b> B. Boston, J. W. Grove, R. Holmes, L. F. Henderson, D. Sharp, Y. Yang and Q. Zhang .....	181
<b>Stable Compact Schemes for Shock Calculations</b> Bernardo Cockburn and Chi-Wang Shu .....	195
<b>Numerical Wind Tunnel Testing of Pressurized Tents</b> Neal E. Blackwell .....	211
<b>Three-Dimensional Finite Elements with Multiple-Quadrature-Points</b> Wing Kam Liu, Yu-Kan Hu and Ted Belytschko. ....	229
<b>Three-Dimensional Finite Element Model Generation and Response Simulation of an Armored Vehicle</b> A. D. Gupta, J. M. Santiago and H. L. Wisniewski. ....	247
<b>The Scaling Laws for Fluid Mixing</b> Qiang Zhang and James Glimm .....	261
<b>Stability Analysis of Stochastic PDE's via Lyapunov Functionals</b> Pao-Liu Chow .....	273

<b>Title</b>	<b>Page</b>
Analysis and Computation of Approximate Solutions to a Simple Model of Shear Band Formation in One and Two Dimensions Donald A. French .....	287
A System for Materials of Korteweg Type in Two Space Variables Harumi Hattori and Dening Li .....	299
Dynamics of A Polymer Molecule Near A Single Streamwise Vortex Joseph D. Myers .....	305
A Massively Parallel Iterative Numerical Algorithm For Immiscible Flow in Naturally Fractured Reservoirs Jim Douglas, Jr, P.J. Leme, Felipe Pereira and Li-Ming Yeh. ....	329
Numerical Solution of Richards' Equation A. J. Silva Neto and R. E. White. ....	349
Numerical Solution of Fluid Flow in Partially Saturated Porous Media A. J. Silva Neto and R. E. White .....	353
Explicit Formal Solution to Generalized Kolmogorov Equation Shing-Tung Yau and Stephen S.T. Yau .....	373
Nonlinear Partial Differential Equations of Interest in Nonlinear Optics M. J. Potasek .....	387
Forced Lattice Vibrations Percy Deift and Thomas Kriecherbauer. ....	405
Asymptotic Model for Deflagration-to-Detonation Transition in Reactive Two-Phase Flow Pedro F. Embid .....	423

<u>Title</u>	<u>Page</u>
Thermionic Emission Form High- $T_c$ Superconductors Richard A. Weiss .....	443
Slow and Ultrafast Dynamical Systems Richard A. Weiss .....	487
Ultrafast Quantum Processes Richard A. Weiss .....	527
Quaternary Fission of $\gamma$ Ray Cooled Actinides Richard A. Weiss .....	577
Target Tracking and Recognition Using Jump- Diffusion Processes Anuj Srivastava, Robert S. Teichman and Michael I. Miller .....	665
Summary of Heavy Traffic Convergence of a Controlled, Multi-Class Queueing System L.F. Martins, S. E. Shreve and H. M. Soner .....	671
Stability and Error Estimates of Stochastic Integro- Differential Equations with Random Parameters G. S. Ladde and S. Sathananthan. ....	691
Numerical Treatment of Random Differential Equations G. S. Ladde and R. Pirapakaran .....	699
Diagonalization and Stability of Two-Time Scale Singularly Perturbed Linear Integro-Differential System G. S. Ladde and S. Sathananthan .....	713
Participant List .....	721

## **ELEVENTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING**

**Host**

**Carnegie Mellon University, Pittsburgh, Pennsylvania**

**8-10 June 1993**

## AGENDA

**Tuesday, 8 June 1993**

**0745 - 1600      Registration - Wean Hall, Room 7500****0815 - 0830      Opening Remarks - Wean Hall, Room 7500****0830 - 0930      General Session I - Wean Hall, Room 7500**

**Chairperson: Benjamin Cummings, U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland**

# COMPLEXITY THEORETIC ASPECTS OF PROBLEMS IN CONTROL THEORY

**John N. Tsitsiklis, Massachusetts Institute of Technology, Cambridge, Massachusetts**

**0930 - 1000 Break**

**1000 - 1200      Special Session 1 - Mathematics for Materials  
- Wean Hall, Room 7500**

**Chairperson: Julian Wu, U.S. Army Research Office,  
Research Triangle Park, North Carolina**

## KINEMATICS OF INCOHERENT PHASE TRANSITIONS

**Mort Gurtin, Carnegie Mellon University,  
Pittsburgh, Pennsylvania**

## METASTABILITY IN PHASE TRANSFORMATIONS

**R. D. James, University of Minnesota,  
Minneapolis, Minnesota**

**Tuesday (continued)**

**THE CRYSTALLINE ALGORITHM FOR COMPUTING MOTION BY CURVATURE**

Robert V. Kohn, New York University, New York, NY

**VISCOSITY SOLUTIONS FOR SIMPLIFIED HYPOPLASTICITY MODELS**

Michael Shearer, North Carolina State University,  
Raleigh, North Carolina

xxxxxxxxxxxxxxxxxxxxxxxxxxxx

1000 - 1200

**Technical Session 1-Algebraic and Symbolic Methods  
-Physical Plant Building, Room 300B**

**Chairperson: William Jackson, U.S. Army Tank-Automotive  
Command, Warren Michigan**

**GENERALIZED EUCLIDEAN ALGORITHMS**

Moss Sweedler, MSI of Cornell University,  
Ithica, New York

**DECODING CASCADED REED-SOLOMON CODES**

Keith Saints, MSI of Cornell University,  
Ithica, New York

**THE LOGIC OF PROGRAM CONTROL: EXTENDING THE CURRY-  
HOWARD ISOMORPHISM**

Judith Underwood, MSI of Cornell University,  
Ithica, New York

**NEWTONIAN GRAPHS AND ALGEBRAIC FUNCTIONS**

Kjartan Stefansson and D. Kozen, MSI of  
Cornell University, Ithica, New York

**THE MODELING AND ANALYSIS OF SOFTWARE PRODUCTION  
PROCESSES: AN APPROACH VIA PETAN**

S. E. Elmaghraby, E. I. Baxter and Mladen A. Vouk,  
North Carolina State University, Raleigh, North Carolina

**Tuesday (continued)**

**IMPLEMENTING MIXED CHAINING IN A CLASSIFICATION TYPE  
EXPERT SYSTEM**

Andrew W. Harrell, U.S. Army Waterways Experiment  
Station, Vicksburg, Mississippi

**1200 - 1330**

**Lunch**

**1330 - 1530**

**Technical Session 2 - Optimization and Nonlinear Dynamics  
- Wean Hall, Room 7500**

**Chairperson: Royce Soanes, Benet Laboratories, Watervliet,  
New York**

**SEVERLY CONSTRAINED ALLOCATION OF BOUNDED RESOURCES**

T. Cronin, Intelligence Electronic Warfare Directorate,  
Warrenton, Virginia

**THEORETICAL ALGORITHMS FOR SOLVING THE ARMY STATIONING  
PROBLEM**

Janet Hurst Spoonamore, US Army Construction  
Engineering Research Laboratory, Champaign, Illinois

**CEMPLOT: A VISUALIZATION TOOL FOR CEM VIII**

K. Shivaswamy, D. Alciatore and P. Burns,  
Colorado State University, Fort Collins, Colorado

**THE TODA LATTICE UNDER PERIODIC FORCING**

P. Delft, T. Kriecherbauer and S. Venakides,  
Duke University, Durham, North Carolina

**DYNAMICAL APPROACH TO COGNITION: THE ROLE OF  
BIFURCATION AND CHAOS**

Nabil H. Farhat, University of Pennsylvania,  
Philadelphia, Pennsylvania

**STEADY STATE SOLUTIONS TO A FORCED NONLINEAR  
OSCILLATOR: A COMPARISON OF RESULTS USING A GENERALIZED  
HARMONIC BALANCE METHOD AND BY NUMERICAL  
INTERGRATION**

Julian Wu and Ben Noble, US Army Research Office,  
Research Triangle Park, North Carolina

**Tuesday (continued)**

xxxxxxxxxxxxxxxxxxxx

**1330 - 1530**

**Technical Session 3 - Computational Methods (Finite Element/  
Finite Difference)  
- Physical Plant Building, Room 300B**

**Chairperson: J. Michael Coyle, Benet Laboratories,  
Watervliet, New York**

**THE ANALYSIS OF SHOCK INDUCED SURFACE INSTABILITIES**

**J. W. Grove, R. Holmes, Y. Yang, Q. Zhang and D. H. Sharp,  
State University of New York at Stony Brook,  
Stony Brook, New York**

**NONLINEARLY STABLE COMPACT SCHEMES FOR SHOCK  
CALCULATIONS**

**Chi-Wang Shu, Brown University, Providence, RI**

**NUMERICAL WINDTUNNEL TESTING OF CP DEPMEDS**

**Neal Blackwell, U.S. Army Belvoir Research,  
Development and Engineering Center,  
Fort Belvoir, Virginia**

**MULTIPLE-QUADRATURE-POINT ALL FINITE ELEMENTS**

**W. K. Liu, Y-K. Hu, T. Belytschko,  
Northwestern University, Evanston, Illinois**

**THREE-DIMENSIONAL FINITE ELEMENT MODEL GENERATION OF  
AN ARMORED VEHICLE FOR RESPONSE SIMULATION**

**A. D. Gupta, J. M. Santiago and H. L. Wisniewski,  
U.S. Army Research Laboratory,  
Aberdeen Proving Ground, Maryland**

**AUTOMATED SYNTHESIS OF PARALLEL FEA PROGRAMS**

**Naveen Sharma, Kent State University, Kent, Ohio**

## Tuesday (continued)

**1530 - 1600**      **Break**

**1600 - 1700                      General Session II - Wean Hall, Room 7500**

**Chairperson: Kenneth D. Clark, U.S. Army Research Office,  
Research Triangle Park, North Carolina**

# NUMERICAL SOLUTIONS IN MICROSTRUCTURE AND THE CALCULUS OF VARIATIONS

**Roy A. Nicolaides, Carnegie Mellon University,  
Pittsburgh, Pennsylvania**

**Wednesday, 9 June 1993**

**0800 - 1400      Registration**

**0830 - 0930      General Session III - Wean Hall, Room 7500**

**Chairperson: Norman Coleman, U.S. Army Armament Research, Development and Engineering Center, Picatinny Arsenal, New Jersey**

## RATIONAL WAVELETS IN CONTROL

**P. S. Krishnaprasad, University of Maryland, College Park, Maryland**

**0930 - 1000**                      **Break**

**1000 - 1200      Special Session II • Topics in Stochastic Analysis  
• Wean Hall, Room 7500**

**Chairperson: Gerald Andersen, U.S. Army Research Office,  
Research Triangle Park, North Carolina**

## SOME RESULTS ON BROWNIAN TOURISTS

**Michael Cranston, Cornell University, Ithaca, NY**

## THE SCALING LAWS FOR FLUID MIXING

**Qiang Zhang, SUNY at Stony Brook, Stony Brook, NY**



**Wednesday (continued)**

**STABILITY ANALYSIS OF STOCHASTIC PDE's VIA LYAPOUNOV  
FUNCTIONALS**

Pao-Liu Chow and R. Khasminskii,  
Wayne State University, Detroit, Michigan

xxxxxxxxxxxxxxxxxxxxxxxx

**1000 - 1200**

**Technical Session 4 - Mathematics of Material Sciences  
- Physical Plant Building, Room 300B**

**Chairperson: John D. Vassilakis, Benet Laboratories,  
Watervliet, New York**

**ON THE NUMERICAL APPROXIMATION OF A MODEL FOR  
SHEARBAND FORMATION**

Donald A. French, University of Cincinnati,  
Cincinnati, Ohio

**A 2-DIMENSIONAL SYSTEM OF MIXED TYPE FOR MATERIALS OF  
KORTEWEG TYPE**

H. Hattori and D. Li, West Virginia University,  
Morgantown, West Virginia

**COMPUTATIONAL RESULTS FOR THE DYNAMICS OF THE  
AUSTENITIC-MARTENSITIC INTERFACE**

Peter Kloucek and Mitchell Luskin,  
University of Minnesota, Minneapolis, Minnesota

**DYNAMICS OF A POLYMER MOLECULE NEAR A SINGLE  
STREAMWISE VORTEX**

F. H. Abernathy and J. D. Myers,  
U.S. Military Academy, West Point, New York

**NUMERICAL TECHNIQUES FOR TREATING PROBLEMS OF  
QUASISTATIC AND DYNAMIC LINEAR VISCOELASTICITY**

S. Shaw, M. Warby and J. R. Whiteman,  
Brunel-The University of West London,  
Uxbridge, Middlesex, United Kingdom

**Wednesday (continued)**

**ANALYSIS OF THICK COMPOSITE SHELLS**

**T. Tsui and E. Saether,  
U.S. Army Research Laboratory, Watertown, Massachusetts**

**1200 - 1330**

**Lunch**

**1330 - 1530**

**Special Session 3 - Computation in Environmental  
Geoscience  
- Wean Hall, Room 7500**

**Chairperson: Kenneth D. Clark, U.S. Army Research office,  
Research Triangle Park, North Carolina**

**DEVELOPMENT OF A GROUNDWATER MODELING SYSTEM FOR  
REMEDICATION OF CONTAMINATED DOD SITES**

**Jeffrey Holland, U.S. Engineering Waterways  
Experiment Station, Vicksburg, Mississippi**

**SIMULATIONS OF TURBULENT MIXING EMPLOYING  
HETEROGENEOUS SUPERCOMPUTING**

**Andrei Malevsky, University of Minnesota,  
Minneapolis, Minnesota**

**A PARALLEL ITERATIVE NUMERICAL ALGORITHM FOR  
IMMISCIBLE FLOW IN NATURALLY FRACTURED RESERVOIRS**

**Felipe Pereira, Purdue University, West Lafayette, IN**

**NUMERICAL SOLUTION OF RICHARDS' EQUATION**

**R. E. White, North Carolina State University,  
Raleigh, North Carolina**

**xxxxxxxxxxxxxxxxxxxxxxxx**

**1330 - 1530**

**Special Session 4 - Virtual Factory  
- Physical Plant Building, Room 300B**

**Chairperson: Benjamin Cumminge, U.S. Army Research  
Laboratory, Aberdeen Proving Ground, MD**

**Wednesday (continued)**

***TITLE NOT AVAILABLE AT THIS TIME***

**Avner Friedman, University of Minnesota,  
Minneapolis, Minnesota**

***TITLE NOT AVAILABLE AT THIS TIME***

**Albert Garcia, Defense Systems Management College,  
Fort Belvoir, Virginia**

**1530 - 1600**

**Break**

**1600 - 1730**

**General Session IV - Wean Hall, Room 7500**

**Chairperson: Jagdish Chandra, U.S. Army Research Office,  
Research Triangle Park, North Carolina**

***MATHEMATICAL COMPUTING IN MODELING 1993***

**Frank R. Giordano, U.S. Military Academy,  
West Point, New York**

***THE ROLE OF HARD THERMAL CONVECTION IN GEOSCIENCES***

**David Yuen, University of Minnesota,  
Minneapolis, Minnesota**

**Thursday, 10 June 1993**

**0730 - 1200**

**Registration - Wean Hall, Room 7500**

**0800 - 1020**

**Technical Session 5 - Mathematical Methods in Physical  
Sciences - Wean Hall, Room 7500**

**Chairperson: Terence Cronin, Intelligence Electronic  
Warfare Directorate, Warrenton, Virginia**

***EXPLICIT ALGORITHM FOR COMPUTING THE FUNDAMENTAL  
SOLUTION TO KOLMOGOROV EQUATION***

**Stephen Yau, The University of Illinois at Chicago,  
Chicago, Illinois**

***AN EXAMPLE OF HOW INSIGHT, CLOSED-FORM MATHEMATICAL  
MODELING AND COMPUTER-BASED SYMBOLIC MATHEMATICS CAN  
HELP IN DOING INTERESTING SCIENCE: NUCLEAR SPIN  
RELAXATION IN SOLIDS***

**P. A. Beckmann, A. M. Albano and C. A. Palmer,  
Bryn Mawr College, Bryn Mawr, Pennsylvania**

**Thursday (continued)**

**NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS OF INTEREST  
IN NONLINEAR OPTICS**

M. J. Potasek, Columbia University in the City of New York,  
New York, New York

**ALGORITHMS FOR THE SCATTERING OF BELTRAMI FIELDS**

B. Shanker and A. Lakhtakia, The Pennsylvania State  
University, University Park, Pennsylvania

**AN ASYMPTOTIC THEORY FOR HOT SPOT FORMATION AND  
TRANSITION TO DETONATION FOR REACTIVE GRANULAR  
MATERIALS**

Pedro F. Embid, University of New Mexico,  
Albuquerque, New Mexico

**A 2-DIMENSIONAL SYSTEM FOR FLUIDIZED BED**

G. Ganser, X. Hu and D. Li, West Virginia University,  
Morgantown, West Virginia

**1) THERMIONIC EMISSION FROM HIGH-TC SUPERCONDUCTORS**

**2) SLOW AND ULTRAFAST DYNAMICAL SYSTEMS**

Richard A. Weiss, U.S. Army Waterways Experiment  
Station, Vicksburg, Mississippi

xxxxxxxxxxxxxxxxxxxxxxxx

**0800 - 1020**

**Technical Session 6 - Systems, Control, and Stochastic  
Processes  
- Physical Plant Building, Room 300B**

**Chairperson: Louis Picotelle, U.S. Army Natick Research  
and Development Command, Natick, MA**

**A COOPERATIVE-COMPETITIVE DIFFERENTIAL GAME WITH THE  
DEVIL WITH APPLICATIONS TO ROBUST AND DECENTRALIZED  
CONTROL**

C. A. Jacobson and Gilead Tadmor,  
Northeastern University, Boston, Massachusetts

**Thursday (continued)**

**MODEL CONVERSIONS OF UNCERTAIN LINEAR SYSTEMS  
USING A SCALE AND SQUARING GEOMETRIC SERIES METHOD**

L. S. Shieh, J. Gu and J. S. Tsai,  
University of Houston, Houston, Texas

**NARROWBAND TRACKING AND TARGET RECOGNITION USING  
JUMP-DIFFUSION PROCESSES**

A. Srivastava, R. S. Teichman and M. I. Miller,  
Washington University, Saint Louis, Missouri

**APPROXIMATION OF COMPLEX CONTROLLED NETWORKS**

H. Mete Soner, Carnegie Mellon University,  
Pittsburgh, Pennsylvania

**ALMOST SURE CONVERGENCE OF STOCHASTIC INTERACTIVE  
PROCESSES**

G. S. Ladde and S. Sathananthan,  
The University of Texas at Arlington, Arlington, Texas

**STABILITY AND ERROR ESTIMATES OF STOCHASTIC INTEGRO  
DIFFERENTIAL EQUATIONS WITH RANDOM PARAMETERS**

G. S. Ladde and S. Sathananthan, Tennessee State  
University, Nashville, Tennessee

**CONVERGENCE AND STABILITY OF LARGE-SCALE STOCHASTIC  
SYSTEMS**

Bonita A. Lawrence and G. S. Ladde,  
The University of Texas at Arlington, Arlington, Texas

**1020 - 1050**

**Break**

**1050 - 1150**

**General Session V - Wean Hall, Room 7500**

**Chairperson: Gerald Andersen, U.S. Army Research Office,  
Research Triangle Park, North Carolina**

**STOCHASTIC SPATIAL MODELS OF EPIDEMICS AND EXCITABLE  
MEDIA**

Richard Durrett, Cornell University, Ithaca, New York

**1150 - 1200**

**Adjournment**

# COMPLEXITY THEORETIC ASPECTS OF PROBLEMS IN CONTROL THEORY<sup>†</sup>

John N. Tsitsiklis  
Center for Intelligent Control Systems  
and  
Laboratory for Information and Decision Systems  
M.I.T.  
Cambridge, MA 02139

**ABSTRACT.** We introduce some of the concepts of computational complexity theory. We then survey available complexity-theoretic results for some of the main problems of control theory.

**INTRODUCTION.** Our subject is motivated by asking what it means "to have a solution to a problem." The answer to this question has been changing with time: for example, ancient Greeks required a constructive procedure using ruler and compass, classical mathematicians required closed form formulas (maybe involving special functions), etc. The theory of computation, as developed during the last fifty years considers a problem to be solved if an algorithm is provided that can compute its solution. There do exist problems that are provably unsolvable, such as the halting problem or Diophantine equations. Even problems that are solvable in principle, as most problems in control theory are, can be of varying difficulty if one considers the required computational resources, such as running time. Complexity theory is the branch of computer science that deals with the classification of problems in terms of their computational difficulty. In this paper, we define some concepts from complexity theory and provide an overview of existing complexity results for several problems in control theory.

**COMPLEXITY THEORY.** Mainstream complexity theory deals with discrete problems; that is, problems whose instances can be encoded in a finite binary string and are therefore suitable input to a digital computer. *Problems* (such as matrix inversion) have *instances* (any particular square matrix defines an instance of matrix inversion). Different instances of the same problem are, in general, of different *sizes*, where "size" means the number of bits used in a natural encoding of that instance. We say that a problem is *polynomial time solvable*, or for short, belongs to  $P$ , if there exists an algorithm and some integer  $k$  such that the time it takes this algorithm to solve any instance of size  $n$  has order of magnitude  $O(n^k)$ . Some classical problems in  $P$  are linear programming, matrix inversion, the shortest path problem in graphs, etc. It is often considered that problems in  $P$  are the "well-solved" ones.

There is another class of problems, called  $NP$  (for nondeterministic polynomial time), that contains all problems that can be "transformed" or reformulated as integer programming problems. While  $P$  is a subset of  $NP$ , there is no known polynomial time algorithm for integer programming and it is generally conjectured that  $P \neq NP$ . If this conjecture is true, then integer programming is not solvable in polynomial time and the same is true for those problems in  $NP$  which are the "hardest"; such problems are called  $NP$ -complete. More generally, we will say that a problem is  $NP$ -hard if it is at least as hard as integer programming.

Proving that a problem is  $NP$ -hard is viewed as evidence that the problem is difficult. Assuming the validity of the conjecture  $P \neq NP$ ,  $NP$ -hard problems do not have polynomial time algorithms. More generally,  $NP$ -hardness often reflects a certain absence of structure which limits the nature of theoretical results that can be established. In practical terms,  $NP$ -hardness usually

---

<sup>†</sup> Research supported by the ARO under grant DAAL03-92-G0309.

means that a problem should be approached differently: instead of trying to develop an algorithm which can be proved to work efficiently all of the time, effort should be concentrated on easier special cases or on heuristics that work acceptably well most of the time; usually, this is to be determined by extensive experimentation rather than by theoretical means.

**DECENTRALIZED DECISION MAKING.** Witsenhausen's problem [W68] is the simplest conceivable generalization of linear quadratic Gaussian control (LQG) to a nonclassical information pattern (decentralized control). The solution to this problem has remained unknown despite persistent efforts. An explanation is provided by the fact that this problem, suitably discretized, is *NP*-hard [PT86]. The point here is not that we might wish to solve Witsenhausen's problem computationally; *NP*-hardness is an indication that the problem is fundamentally more difficult — and less structured — than its centralized analog (LQG).

The above result is not fully satisfactory because it does not rule out the possibility that *NP*-hardness is only a consequence of the problem discretization, not of the inherent difficulty of the original problem. While it seems difficult to establish a complexity result for the original (continuous) version of the problem, results similar to *NP*-hardness have been established for a related problem which we discuss below.

In the *team decision* problem [MR72], we are given two random variables  $y_1, y_2$ , with known joint probability distribution, and a cost function  $c : \mathbb{R}^4 \mapsto \mathbb{R}$ . Agent  $i$  ( $i = 1, 2$ ) observes the value of  $y_i$  and makes a decision  $u_i$  according to a rule  $u_i = \gamma_i(y_i)$ . A cost of  $c(y_1, y_2, u_1, u_2)$  is then incurred. The problem is to find rules  $\gamma_1$  and  $\gamma_2$ , so as to minimize the expectation of the cost. This problem is *NP*-hard for the case where the  $y_i$ s have finite range [PT82, TA85]. It remains *NP*-hard even for a special case that arises in decentralized detection [TA85].

In the continuous version of the problem, we take the random variables  $y_1$  and  $y_2$  to be uniformly distributed on  $[0, 1]$ . The function  $c$  is assumed to be Lipschitz continuous with known Lipschitz constant. Such a function cannot be represented by a binary string, as required by digital computers, and for this reason, we need a suitably adapted model of computation: we assume that a digital computer obtains information on the function  $c$  by submitting queries to an "oracle"; a typical query consists of a rational (hence finitely describable) vector in the domain of  $c$ , together with an integer  $k$ ; the oracle returns the  $k$  most significant bits of the value of  $c$  at that point. Finally, there is an accuracy parameter  $\epsilon$ : instead of looking for an optimal solution, we only desire a solution whose expected cost comes within  $\epsilon$  of the optimal. The "cost" or running time under this model of computation can be viewed as consisting of two parts: a) the number of queries (information cost) and b) the time spent on actual computations (combinatorial cost).

For the continuous version of the team decision problem, the cost of information is only  $O(1/\epsilon^4)$ ; it suffices to learn the value of the function  $c$  at points on a grid with spacing  $\epsilon$  and any smaller number of queries is insufficient. On the other hand, assuming that  $P \neq NP$ , there exists no algorithm that solves the problem with accuracy  $\epsilon$  in time polynomial in  $1/\epsilon$  [PT86].

Two remarks are in order:

- a) For many problems whose complexity has been studied within this framework, the information cost and the combinatorial cost are comparable. Examples can be found in numerical integration, numerical integration of PDEs, nonlinear programming, etc. In contrast, we have here an exponential gap between the two types of costs.
- b) Problems with closed form solutions often have complexity which is polynomial in the logarithm of  $1/\epsilon$ . Even problems like PDEs have complexity of the order of  $1/\epsilon^k$ , where the exponent  $k$  depends on the dimension of the problem and the smoothness of the data. In this respect, the team decision problem is significantly harder than most PDE problems.

## MARKOV DECISION THEORY

Consider a controlled Markov chain  $x(t)$  that takes values in the finite set  $\{1, \dots, n\}$ . We are given its transition probabilities  $p(x(t+1) | x(t), u(t))$ , where  $u(t)$  is the control applied at time  $t$ . The cost per stage is of the form  $c(x(t), u(t))$  and there is a discount factor  $\alpha \in (0, 1)$ . The objective is to minimize the infinite horizon expected cost

$$V(i) = E\left[\sum_{t=0}^{\infty} \alpha^t c(x(t), u(t)) \mid x(0) = i\right]$$

over all causal control policies. The key to solving this problem is Bellman's equation:

$$V(i) = \min_u [c(i, u) + \alpha \sum_j p(j | i, u) V(j)]$$

While there are several algorithms for solving Bellman's equation (e.g., policy iteration, successive approximation, etc.), none of the algorithms of this type is known to run in polynomial time. On the other hand, this problem is polynomially solvable because it can be transformed to linear programming [B87].

The problem becomes more difficult if the control is to be chosen on the basis of imperfect information. Suppose that at each time  $t$ , we observe  $y(t) = h(x(t))$ , where  $h$  is a known function. We restrict to policies in which the current decision  $u(t)$  is determined by a rule of the form  $u(t) = F(y(0), y(1), \dots, y(t), t)$ . If we let  $p(i, t) = \Pr(x(t) = i \mid \text{past history})$ , the problem can be reformulated as a perfect information problem with state vector  $p(t) = (p(1, t), \dots, p(n, t))$ . The cost-to-go function  $V(p(t))$  is known to be piecewise linear in  $p$  and this leads to a finite algorithm [SS73], at least for the case of finite-horizon problems. Unfortunately, the number of linear "faces" can increase exponentially with the time horizon, and so does the required algorithm. Are more efficient algorithms likely to exist? The answer is probably negative because the problem is NP-hard [PT87].

There are analogs of these results that apply to the problem of supervisory control of discrete-event systems, as formulated by Ramadge and Wonham [RW87]. These problems are similar to the problems of Markov decision theory except that the transition probabilities are not given (or may not exist) and the problem consists of finding feedback laws that are guaranteed to avoid certain undesirable states. While the perfect information problem was known to be polynomial, the corresponding imperfect information problem, as well as the corresponding problem of decentralized control, are NP-hard [T88].

The Markov decision problem has also been extensively studied for the case of continuous state spaces. In a simple version of the problem, we may assume that the state and the control take values in  $[0, 1]$ . Bellman's equation becomes

$$V(x) = \min_u [c(x, u) + \alpha \int_0^1 p(y | x, u) V(y) dy]$$

Let us assume that the functions  $c$  and  $p$  have bounded first derivatives. If we wish to solve Bellman's equation with accuracy  $\epsilon$ , it is not hard to show that  $O(1/\epsilon^3)$  "oracle queries" suffice; it turns out that this many queries are also necessary [CT89]. The natural iterative method for this problem (successive approximation) has computational complexity  $O((1/\epsilon^3) \log(1/\epsilon))$ : the cost per iteration is  $O(1/\epsilon^3)$  and  $\log(1/\epsilon)$  iterations are needed to get within  $\epsilon$  of the solution. In fact, the logarithmic gap between the lower bound of  $O(1/\epsilon^3)$  and the performance of successive approximation can be closed. It turns out that a "one-way multigrid" method solves the problem with a total of  $O(1/\epsilon^3)$  arithmetic operations and is therefore an optimal algorithm [CT91]. The key idea is that most iterations are performed on coarse grids and are therefore relatively inexpensive.



**OTHER RESULTS.** We mention briefly some more problems in control theory for which complexity theoretic results are available.

**Nonlinear controllability:** The question here is whether it is possible to generalize Kalman's controllability conditions to nonlinear systems. Consider a bilinear system of the form

$$\frac{dx}{dt} = (A + \sum_{i=1}^m u_i G_i)x + Bu, \quad x \in M,$$

where  $M$  is a manifold defined as the zero set of a given set of polynomials. It turns out that deciding whether such a system is controllable (i.e., whether every state can be reached from any other state) is  $NP$ -hard [S88]. As a result, whatever necessary and sufficient conditions for controllability are found will be computationally very difficult to test (unless  $P = NP$ ).

**Robust control:** Consider the linear system  $dx/dt = Az$  and assume that  $A = A_0 + \alpha_1 A_1 + \dots + \alpha_k A_k$ , where each  $\alpha_i$  is an unknown parameter that lies in  $[0, 1]$ . We are interested in properties of the plant that are valid for all choices of  $\alpha_1, \dots, \alpha_k$ . It turns out that many interesting problems within this framework are unlikely to have computationally efficient solutions. For example, deciding whether  $A$  is guaranteed to be nonsingular or stable is  $NP$ -hard [PR92,N92]. As a corollary of this result, computing the structured singular value  $\mu$  of a linear system is also an  $NP$ -hard problem [BYDM92].

**Simultaneous stabilization:** Let there be given matrices  $A_1, \dots, A_k$  and  $B$ . The problem is whether there exists a matrix  $K$  such that  $A_i - BK$  is stable for all  $i$ . This problem can be also shown to be  $NP$ -hard.

**Control of queueing systems:** Consider the standard problem of controlling a closed queueing network with several servers and several customer classes. (Control involves routing decisions for customers that complete service and sequencing decisions; the latter deal with choosing which customer to serve next at any given server with several customers in queue.) The objective is to find a control policy that minimizes the expected weighted sum of queue lengths. This problem is  $NP$ -hard even in the special case where the service time distributions are exponential [PT93]. In addition, its computational complexity is exponential for the case of deterministic service times [PT93]. Note that the latter result is stronger than any of the results mentioned earlier in this paper in that it does not rely on the validity of the conjecture  $P \neq NP$ .

**CLOSING COMMENTS.** Complexity theory can provide new insights to control problems. However, any results obtained have to be interpreted with caution. Proving that a problem is  $NP$ -hard does not mean that the problem is intractable and that work on it should be abandoned. Rather, a different line of attack may be called for.

## REFERENCES

- [B87] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Prentice Hall, 1987.
- BYDM92] R. D. Braatz, P. M. Young, J. C. Doyle and M. Morari, "Computational Complexity of  $\mu$  Calculation", preprint, 1992.
- [CT89] C.-S. Chow and J.N. Tsitsiklis, "The Complexity of Dynamic Programming", *Journal of Complexity*, Vol. 5, 1989, pp. 466-488.
- [CT91] C.-S. Chow and J.N. Tsitsiklis, "An Optimal One-Way Multigrid Algorithm for Discrete-Time Stochastic Control", *IEEE Transactions on Automatic Control*, Vol. AC-36, No. 8, 1991, pp. 898-914.
- [MR72] J. Marschak and R. Radner, *The Economic Theory of Teams*, Yale Univ. Press, 1972.
- [N92] A. S. Nemirovsky, "Several  $NP$ -hard Problems Arising in Robust Stability Analysis", preprint, 1992.

- [PT82] C.H. Papadimitriou and J.N. Tsitsiklis, "On the Complexity of Designing Distributed Protocols", *Information and Control*, Vol. 53, No. 3, 1982, pp. 211-218.
- [PT86] C.H. Papadimitriou and J.N. Tsitsiklis, "Intractable Problems in Control Theory", *SIAM Journal on Control and Optimization*, Vol. 24, No. 4, 1986, pp. 639-654.
- [PT87] C.H. Papadimitriou and J.N. Tsitsiklis, "The Complexity of Markov Decision Processes", *Mathematics of Operations Research*, Vol. 12, No. 3, 1987, pp. 441-450.
- [PT93] C.H. Papadimitriou and J.N. Tsitsiklis, "The Complexity of Networks of Queues," in preparation.
- [PR92] S. Poljak and J. Rohn, "Checking Robust Nonsingularity is NP-hard", preprint.
- [RW87] P. J. Ramadge and W. M. Wonham, "Supervisory Control of a Class of Discrete-Event Processes", *SIAM J. Control and Optimization*, Vol. 25, 1987, pp. 206-230.
- [S88] E. Sontag, "Controllability is Harder to Decide than Accessibility" *SIAM J. Control and Optimization*, Vol. 26, 1988, pp. 1106-1118.
- [SS73] R. D. Smallwood and E. J. Sondik, "The Optimal Control of Partially Observable Markov Decision Processes over a Finite Horizon", *Operations Research*, Vol. 11, 1973, pp. 1971-1088.
- [T88] J.N. Tsitsiklis, "On the Control of Discrete Event Dynamical Systems", *Mathematics of Control, Signals and Systems*, Vol. 2, No. 2, 1989, pp. 95-107.
- [TA85] J.N. Tsitsiklis and M. Athans, "On the Complexity of Decentralized Decision Making and Detection Problems", *IEEE Transactions on Automatic Control*, Vol. 30, No.5, 1985, pp. 440-446.
- [W68] H. S. Witsenhausen, "A Counterexample in Stochastic Optimum Control," *SIAM J. Control and Optimization*, Vol. 6, pp. 138-147.

# INCOHERENT SOLID-SOLID PHASE TRANSITIONS<sup>1</sup>

Paolo Cermelli

Dipartimento di Matematica

Universita' di Torino, Torino 10123, Italy

Morton E. Gurtin

Department of Mathematics

Carnegie Mellon University, Pittsburgh, PA 15213, USA

## ABSTRACT

Incoherent phase transitions are more difficult to treat than their coherent counterparts. The interface, which appears as a single surface in the deformed configuration, is represented in its undeformed state by a separate surface in each phase. This leads to a rich but detailed kinematics, one in which defects such as vacancies and dislocations are generated by the moving interface. In this paper we discuss incoherent phase transitions in the presence of deformation and mass transport, neglecting inertia. The phase interface is presumed sharp and structured by energy and stress. The final results are a complete set of interface conditions for an evolving incoherent interface.

## KINEMATICS

In a coherent phase transition the body  $B$  occupies a fixed region of space in a uniform reference configuration, the individual phases, which we label  $i=1,2$ , occupy complementary subregions  $B_i(t)$  of  $B$ , and motions are continuous across the undeformed phase interface  $S(t) = B_1(t) \cap B_2(t)$ . As is clear from the statical treatments of Cahn and Larché [1982], Larché and Cahn [1985], and Leo and Sekerka [1989], incoherent phase transitions are far more complicated. The interface, which appears as a single surface in the deformed body, is represented in its undeformed state by a *separate surface*  $S_i(t)$  for each phase  $i$ , even though we choose uniform reference

<sup>1</sup>Supported by the U. S. Army Research Office. This paper presents a synopsis of Cermelli and Gurtin [1994a,1994b]

configurations for the two phases with corresponding reference lattices coincident. Such complications lead to a rich but detailed kinematics, one in which defects such as dislocations, vacancies, and interstitials may be generated by the moving interface.<sup>2</sup>

A two-phase motion is a pair  $\mathbf{y}=(\mathbf{y}_1, \mathbf{y}_2)$ : at each time  $t$ ,  $\mathbf{y}_i$  maps material points  $\mathbf{X}$  in the undeformed region  $B_i$  for phase  $i$  into points  $\mathbf{x}=\mathbf{y}_i(\mathbf{X}, t)$  in the deformed body.  $\mathbf{F}$  is then the deformation gradient:  $\mathbf{F}=\nabla \mathbf{y}_1$  in phase 1,  $\mathbf{F}=\nabla \mathbf{y}_2$  in phase 2; in addition, we denote by  $\mathbf{F}_i$  the limit of  $\mathbf{F}$  as the interface is approached from phase  $i$ . Associated with each two-phase motion are three basic kinematical quantities:

- (1) The *incoherency tensor*  $\mathbf{H}$ , which measures the stretching and twisting of one phase relative to the other at the interface.  $\mathbf{H}$  is the tangential part of the relative deformation gradient

$$\mathbf{H} = \mathbf{F}_2^{-1} \mathbf{F}_1. \quad (1)$$

For any point  $\mathbf{x}$  of the deformed interface,  $\mathbf{H}$  is a linear transformation  $d\mathbf{X}_2 = \mathbf{H} d\mathbf{X}_1$  between infinitesimal line segments  $d\mathbf{X}_i$  on  $S_i$  that coincide at  $\mathbf{x}$  when deformed. If, for all such line segments,  $d\mathbf{X}_2 = d\mathbf{X}_1$  (or  $d\mathbf{X}_2 = \mathbf{Q} d\mathbf{X}_1$  with  $\mathbf{Q}$  a symmetry rotation of the lattice), then the deformed lattices fit together and the interface is *infinitesimally coherent* at  $\mathbf{x}$ .

- (2) The *production-rate of lattice points*, as measured by the jump  $[W]$  in the interfacial volume flows  $W_i = V_i / \mathcal{J}_i$ , where, for each  $i$ ,  $V_i$  is the normal velocity of  $S_i$ , while  $\mathcal{J}_i$  is the surface Jacobian for  $\mathbf{y}_i$  considered as a deformation of  $S_i$ .
- (3) The *slip*, as measured by the difference  $(\mathbf{y}_2)^\circ - (\mathbf{y}_1)^\circ$ , where  $(\mathbf{y}_i)^\circ$  is the time derivative of  $\mathbf{y}_i$  following the normal trajectories of  $S_i(t)$ .

The incoherency tensor, the lattice-point production, and the slip completely characterize incoherency: *an initially coherent motion is coherent for all time if and only if, at each time, the interface is infinitesimally coherent and the slip and lattice-point production vanish identically.*

<sup>2</sup>Dislocations are discussed by Brooks [1952], Nye [1953], Frank [1955], Bilby [1955], Bilby, Bullough, and De Grinberg [1964], Christian [1965, 1985], Bollman [1967], Christian and Crocker [1980], Pond [1985, 1989], Christian and Crocker [1980], p. 181 and Larché and Cahn [1985], p. 1587 note the possibility of vacancies and dislocations.

## THEORY WITHOUT INTERFACIAL STRUCTURE

The basic physical principles upon which the theory is based are balance of forces, balance of mass, and a version of the second law of thermodynamics appropriate to a mechanical theory.<sup>3</sup> The standard forces associated with continua arise as a response to the motion of material points. The mechanical description of a phase transition requires additional forces<sup>4</sup> that act in response to microstructural changes at the phase interface. We refer to the former as *deformational forces*, to the latter as *configurational forces*.<sup>5</sup> What is most important is that, in addition to the usual force and moment balances for deformational forces, we postulate an additional balance for configurational forces.

We assume that there are  $\mathfrak{A}$  species,  $\alpha = 1, 2, \dots, \mathfrak{A}$ , of mobile atoms with molar densities  $\rho^\alpha$  and corresponding diffusive mass fluxes  $h^\alpha$ .<sup>6</sup> Bulk fields that strongly influence the motion of the interface are the grand canonical potential  $\omega$  and the Eshelby tensor  $P$  defined by

$$\omega = \Psi - \sum_{\alpha=1}^{\mathfrak{A}} \rho^\alpha \mu^\alpha, \quad P = \omega \mathbf{1} - \mathbf{F}^T \mathbf{S}, \quad (2)$$

with  $\Psi$  the bulk energy,  $\mu^\alpha$  the chemical potential of species  $\alpha$ ,  $\mathbf{S}$  the bulk stress, measured per unit undeformed area (Piola-Kirchhoff stress), and  $\mathbf{1}$  the unit tensor.

The final bulk relations are the balance laws

$$\text{Div } \mathbf{S} = 0, \quad (\rho^\alpha)^* = -\text{Div } h^\alpha, \quad (3)$$

supplemented by constitutive equations

$$\begin{aligned} \Psi &= \hat{\Psi}_i(\mathbf{F}, \rho), \quad \mathbf{S} = \partial_{\mathbf{F}} \hat{\Psi}_i(\mathbf{F}, \rho), \quad \mu = \partial_{\rho} \hat{\Psi}_i(\mathbf{F}, \rho), \\ \mathfrak{h} &= -D_i(\mathbf{F}, \rho) \nabla \mu, \end{aligned} \quad (4)$$

<sup>3</sup>Cf. Gurtin [1991].

<sup>4</sup>Cf. the discussion given in the Introduction of Gurtin [1990].

<sup>5</sup>We depart from terminology introduced in Gurtin and Struthers [1990] and Gurtin [1993], where the term *accretive forces* was used.

<sup>6</sup>Cf. Gurtin and Voorhees [1993], Gurtin [1993].

for each phase  $i$ , with

$$\rho = (\rho^1, \dots, \rho^M), \quad \mu = (\mu^1, \dots, \mu^M), \quad \mathbf{h} = (\mathbf{h}^1, \dots, \mathbf{h}^M).$$

We next consider appropriate interface conditions. To best illustrate the basic ideas, we begin with a theory that neglects interfacial energy and stress, but includes interface kinetics. The resulting interface conditions consist of an equation

$$[\mathbf{y}'] \cdot \bar{\mathbf{n}} = -[JW] \quad (5)$$

expressing kinematical compatibility at the interface, a jump condition

$$[\mathcal{J}^{-1} \mathbf{S} \mathbf{n}] = 0 \quad (6)$$

balancing forces across the interface, equations

$$\delta_i \mathbf{n}_i \cdot \mathbf{P}_i \mathbf{n}_i = (\beta_{i1} W_1 + \beta_{i2} W_2) \quad (7)$$

( $i=1,2$ ) balancing normal configurational forces on each phase at the interface, equations

$$(\mathbf{F}_i^T \mathbf{S}_i \mathbf{n}_i)_{\tan S_i} = 0 \quad (8)$$

( $i=1,2$ ) characterizing the vanishing of the tangential traction in each phase at the interface, a relation

$$[\rho^a W] = [\mathcal{J}^{-1} \mathbf{h}^a \cdot \mathbf{n}] \quad (9)$$

expressing mass balance for each species  $a$ , and a condition of local equilibrium

$$[\mu^a] = 0 \quad (10)$$

for each species  $a$ . Here  $\bar{\mathbf{n}}$  is the unit normal to the deformed interface  $\mathcal{S}$ ,  $\mathbf{n}_i$  is the unit normal to the undeformed phase  $i$  interface  $S_i$ ,  $[f]$  denotes

the jump in a bulk field  $f$  across the interface,  $f_i$  denotes the interfacial limit of  $f$  from phase  $i$ , and  $\beta_{ij}$  are kinetic coefficients.

In the derivation of these interface conditions the slip was not included among the independent constitutive variables, a direct consequence of this assumption is (8). The local equilibrium condition (10) is an assumption made from the outset.<sup>7</sup>

The balances (6)-(8) can be expressed more succinctly as a normal force balance

$$[\gamma^{-1}S_n] \cdot \bar{n} = 0 \quad (11)$$

and a partial balance

$$\delta_i P_i n_i = (\beta_{i1} W_1 + \beta_{i2} W_2) n_i \quad (12)$$

for each phase  $i$ .

## THEORY WITH INTERFACIAL STRUCTURE

We turn next to a theory that includes interfacial energy and stress, but neglects mass flow within the interface. Following Cahn and Larché [1982], we choose one of the phases, phase 1, as a reference for the interface, and measure interfacial fields relative to  $S_1$ . Here it is convenient to refer to phase 1 as the *parent phase* and to phase 2 as the *product phase*, and to use the abbreviations

$$S = S_1, \quad n = n_1.$$

We consider a single interfacial energy  $\psi$ , but endow the interface with three stress fields:

- a deformational stress  $\mathbf{S}$  that represents the (Piola-Kirchhoff) stress in the surface and acts in response to the stretching of the parent interface;
- a configurational stress  $\mathbf{C}$  that represents microstructural forces in the parent interface;
- a configurational stress  $\mathbf{K}$  that acts in response to the stretching and rotation of the product-phase lattice relative to the parent-phase lattice.

<sup>7</sup>Cf. Gurtin and Voorhees [1994], who develop a theory in which this assumption is dropped. Their theory neglects deformation.

A consequence of thermodynamics is that the tangential part of the total surface stress

$$\mathbf{A} = \mathbf{C} + \mathbf{F}_1^T \mathbf{S} + \mathbf{H}^T \mathbf{K}, \quad (13)$$

which represents the net configurational contribution of the stresses to the rate of working, is a surface tension whose value is the interfacial energy  $\psi$ .

Among the constitutive equations considered for the interface are relations giving the interfacial energy  $\psi$ , the surface stresses  $\mathbf{S}$  and  $\mathbf{K}$ , and the normal part  $\mathbf{a} = \mathbf{A}^T \mathbf{n}$  of the total surface stress as functions of the limiting value  $\mathbf{F} = \mathbf{F}_1$  of the deformation gradient, the limiting values  $\rho_1$  and  $\rho_2$  of the list of densities, the normal  $\mathbf{n}$  to  $S = S_1$ , and the volume flows  $W_1$  and  $W_2$ . Consequence of the second law are that  $\psi$ ,  $\mathbf{S}$ ,  $\mathbf{K}$ , and  $\mathbf{a}$  are independent of  $\rho_1$ ,  $\rho_2$ ,  $W_1$ , and  $W_2$ ; and that the energy

$$\psi = \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}) \quad (14)$$

generates the stresses through the relations

$$\mathbf{S} = \partial_{\mathbf{F}} \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}), \quad \mathbf{K} = \partial_{\mathbf{H}} \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}), \quad \mathbf{a} = -\partial_{\mathbf{n}} \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}). \quad (15)$$

We show further that  $\psi$ ,  $\mathbf{S}$ , and  $\mathbf{K}$  depend on  $\mathbf{F}$  and  $\mathbf{H}$  through the tangential deformation gradient  $\mathbf{F}$  and the incoherency tensor  $\mathbf{H}$ , that

$$\mathbf{S} = \partial_{\mathbf{F}} \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}), \quad \mathbf{K} = \partial_{\mathbf{H}} \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}), \quad (16)$$

and that  $\mathbf{c} = \mathbf{C}^T \mathbf{n}$  is given by

$$\mathbf{c} = -D_{\mathbf{n}} \hat{\psi}(\mathbf{F}, \mathbf{H}, \mathbf{n}), \quad (17)$$

with  $D_{\mathbf{n}}$  the derivative following  $\mathbf{n}$ .

The final results — which form a complete set of conditions for an incoherent interface — consist of the compatibility condition (5), the mass balance (9), the local equilibrium condition (10), an equation



$$\psi K - (\mathbf{F}^T \mathbf{S} + \mathbf{H}^T \mathbf{K}) \cdot \mathbf{L} + \text{Div}_S \mathbf{C} - \mathbf{n} \cdot \mathbf{P}_1 \mathbf{n} = \beta_{11} W_1 + \beta_{12} W_2 \quad (18)$$

that represents a normal configurational balance for phase 1, an equation

$$\text{Div}_S \mathbf{K} + \mathcal{K} \mathbf{P}_2 \mathbf{n}_2 = \mathcal{K} (\beta_{21} W_1 + \beta_{22} W_2) \mathbf{n}_2 \quad (19)$$

that represents a configurational balance for phase 2, a deformational force balance

$$\text{Div}_S \mathbf{S} + \mathcal{K} \mathbf{S}_2 \mathbf{n}_2 - \mathbf{S}_1 \mathbf{n}_1 = 0, \quad (20)$$

and the constitutive relations (16) and (17). Here  $\mathcal{K} = \mathcal{K}_1 / \mathcal{K}_2$ , while  $\mathbf{L} = \mathbf{L}_1 = -\nabla_S \mathbf{n}$  and  $K = K_1 = \text{tr} \mathbf{L}$ , respectively, are the curvature tensor and the total (twice the mean) curvature for  $S$ .

#### REFERENCES.

- [1952] Brooks, H., Theory of internal boundaries, in: *Metal Interfaces*, Am. Soc. Metals, Cleveland Press 20-64
- [1953] Nye, J. F., Some geometrical relations in dislocated crystals, *Act. Metall.* **1**, 153-162
- [1955] Bilby, B. A., Types of dislocation sources, *Conference on Defects in Crystalline Solids, U. Bristol, Physical Soc. Lond.*, 124-133
- [1955] Frank, F. C., The resultant content of dislocations in an arbitrary intercrystalline boundary, *Symposium on the Plastic Deformations of Crystalline Solids*, Carnegie Institute of Technology, Pittsburgh, 150-154
- [1964] Bilby, B.A., R. Bullough and D. K. De Grinberg, General theory of surface dislocations, in *Dislocations in Solids*, Discussions Faraday Soc., 61-68
- [1965] Christian, J. W., *The Theory of Transformations in Metals and Alloys*, Pergamon Press, Oxford
- [1967] Bollman, W., On the geometry of grain and phase boundaries. 2. Application of the general theory, *Phil. Mag.* **140**, 383-399
- [1980] Christian, J. W. and A. G. Crocker, Dislocations and lattice transformations, in *Dislocations in Solids* (ed. F. Nabarro), North

- Holland, Amsterdam, **3**, 167-249
- [1982] Cahn, J. W. & F. C. Larché, Surface stress and the chemical equilibrium of small crystals. 2. Solid particles embedded in a solid matrix, *Act. Metall.* **30**, 51-56.
- [1985] Christian, J. W., Dislocations and phase transformations, in *Dislocations and Properties of Real Materials*, Inst. Metals, Lond., 94-124
- [1985] Larché, F. C. & J. W. Cahn, The interaction of composition and stress in crystalline solids, *Act. Metall.* **33**, 331-357.
- [1985] Pond, R. C., Interfaces and dislocations, in *Dislocations and Properties of Real Materials*, Inst. Metals, Lond., 71-93
- [1989] Leo, P. H. & R. F. Sekerka, The effect of surface stress on crystal-melt and crystal-crystal equilibrium, *Act. Metall.* **37**, 3119-3138.
- [1989] Pond, R. C., Line defects in interfaces, in *Dislocations in Solids* (ed. F. Nabarro), **8**, North Holland, Amsterdam, 1-65
- [1990] Gurtin, M. E. & A. Struthers, Multiphase thermomechanics with interfacial structure. 3. Evolving phase boundaries in the presence of bulk deformation, *Arch. Rational Mech. Anal.* **112**, 97-160.
- [1991] Gurtin, M. E., On thermodynamical laws for the motion of a phase interface, *Zeit. angew. Math. Phys.*, **42**, 370-388
- [1993] Gurtin, M. E., The dynamics of solid-solid phase transitions. 1. Coherent interfaces, *Arch. Rational Mech. Anal.* Forthcoming
- [1993] Gurtin, M. E. & P. W. Voorhees, The continuum mechanics of coherent two-phase elastic solids with mass transport, *Proc. Roy. Soc. Lond. A*, **440**, 323-343
- [1994a] Cermelli, P. and M. E. Gurtin, On the kinematics of incoherent phase transitions, *Act. Metall.* Submitted
- [1994b] Cermelli, P. and M. E. Gurtin, The dynamics of solid-solid phase transitions. 2. Incoherent interfaces, *Arch. Rational Mech. Anal.* Forthcoming
- [1994] Gurtin, M. E. & P. W. Voorhees, The thermodynamics of nonequilibrium interfaces. 1. General theory. Forthcoming

# Formation and Development of Shear Bands in Granular Material

by

Michael Shearer,<sup>1</sup> F. Xabier Garaizar<sup>2</sup>

Department of Mathematics  
North Carolina State University  
Raleigh, NC 27695

David G. Schaeffer<sup>3</sup> and John Trangenstein<sup>4</sup>

Department of Mathematics  
Duke University  
Durham, NC 27706

## Abstract

A system of equations modeling antiplane shear in a granular material is considered. The model includes the possibility of localization of strain, and the subsequent development of shear bands. This behavior is captured in our analysis of the Riemann initial value problem, in which an initial discontinuity propagates as a combination of moving waves and a stationary shear band. The analytic solution is used to test numerical simulations based on Godunov's method with front tracking and adaptive mesh refinement.

## 1. Introduction

We consider a model [4] for dynamic deformations of granular materials which allows for the localization of flow and the consequent development of shear bands. We focus on Riemann initial value problems in one space dimension that include a shear band. An unusual feature of the solutions is that they are not scale invariant. The structure of the solution (shown in Figure 1) includes the feature that the material unloads elastically between the shear band and a free boundary that propagates into a region of plastic deformation. Mathematically, the Riemann problem reduces to solving a free boundary problem for the (linear) wave equation. We summarize short time existence results and long time behavior, the details of which are given in a series of papers [2, 5]. Then we outline the governing principles in designing an efficient

---

<sup>1</sup>Research supported by NSF grant DMS 9201115, which includes funds from AFOSR, and by ARO grant DAAL03-91-G-0122.

<sup>2</sup>Research supported by NSF grant DMS 9201115, which includes funds from AFOSR.

<sup>3</sup>Research supported by NSF grant DMS 9201034, which includes funds from AFOSR.

<sup>4</sup>Research supported by NSF grant SES-DMS-9201361, by DNA grant DNA001-92-C-0166, and by DOE grant DE-FG05-92ER25145

numerical simulation code, and present numerical results. The numerical results are in excellent agreement with the predictions of the analysis. The simulation code is designed so that it can generalize to higher dimensions; it is currently being extended to two dimensions.

The following system of equations describes antiplane shear deformations that depend on one space variable  $x$  and time  $t$  [4].

$$\begin{aligned} \partial_t v &= \partial_x \sigma & (a) \\ \left[ I + \frac{1}{h(\gamma)} (R T) T^\top \right] \partial_t T &= c^2 \begin{pmatrix} \partial_x v \\ 0 \end{pmatrix} \text{ (if loading)} & (b) \\ \partial_t T &= c^2 \begin{pmatrix} \partial_x v \\ 0 \end{pmatrix} \text{ (otherwise).} & (c) \end{aligned} \quad (1.1)$$

The dependent variables  $v$  and  $T = (\sigma, \tau)^\top$  represent velocity and stress respectively, all other entries of the full stress tensor being constant.  $T^\top$  is the transpose of  $T$ . The constant  $c$  is the elastic wave speed,  $I$  is the  $2 \times 2$  identity matrix and  $R$  is the rotation matrix

$$R = \begin{pmatrix} \cos \beta & \sin \beta \\ -\sin \beta & \cos \beta \end{pmatrix}, \quad (1.2)$$

with parameter  $\beta \in (0, \frac{\pi}{2})$  measuring the degree of nonassociativity in the model (specifically in the flow rule). The given function  $h = h(\gamma)$  is the hardening modulus, depending on the yield stress  $\gamma$ . The function  $h$  is nonnegative and monotonically decreasing on  $[0, 1]$ , with  $h(1) = 0$ .

Equation (1.1a) is conservation of momentum, while equations (1.1b,c) specify constitutive behavior. This behavior is described as loading (or *plastic*) when  $|T|$  is at its maximum over previous time and is increasing. That is, the material is loading when it is at plastic yield:

$$|T(x, t)| = \gamma(x, t) \equiv \max_{0 \leq s \leq t} |T(x, s)|, \quad (1.3)$$

and the right hand side of (1.3) is increasing:

$$\partial_t \gamma > 0. \quad (1.4)$$

When  $|T(x, t)| < \gamma(x, t)$ , the behavior is described as elastic. When (1.3) holds and  $\partial_t \gamma = 0$ , we use the terminology neutral loading.

We have a differential equation for  $\gamma$  :

$$\partial_t \gamma = \begin{cases} \partial_t |T| & \text{if loading} \\ 0 & \text{otherwise} \end{cases} \quad (1.5)$$

The Riemann problem for system (1.1) is the initial value problem with initial data of the form

$$(v, T, \gamma)(x, 0) = \begin{cases} (v_L, T_L, \gamma_L) & \text{if } x < 0 \\ (v_R, T_R, \gamma_R) & \text{if } x > 0, \end{cases} \quad (1.6)$$

subject to  $|T(x, 0)| \leq \gamma(x, 0)$ . In solving Riemann problems for a simplified version of system (1.1), Garaizar [1] found initial data (1.6) for which there is no scale invariant solution. The reason is that equations (1.1) lose hyperbolicity as  $\gamma$  approaches a critical value  $\gamma = \gamma^*$ , and the classical construction of scale invariant solutions breaks down.

When the equations lose hyperbolicity, they also lose linear well-posedness, and we suppose that a shear band forms. As in [4], we treat the shear band as a stationary discontinuity, with nonstandard jump conditions. Specifically, the velocity gradient is approximated by a divided difference:

$$v_x \approx [v]/\delta, \quad (1.7)$$

where  $[v]$  is the jump in velocity across the shear band and  $\delta$  is a small parameter to the grain size. From the conservation of momentum equation (1.1a), we see that  $\sigma$  is continuous across a shear band. The variable  $\tau$  experiences a jump on each side of the shear band. The approximation (1.7) leads to the following system of ordinary differential equations, with a constraint, for evolution of the variables  $T = (\sigma, \tau)^\top, \gamma$  within the shear band.

$$\left[ I + \frac{1}{h(\gamma)} (R \ T) T^\top \right] \partial_t T = c^2 \begin{pmatrix} [v]/\delta \\ 0 \end{pmatrix}, \quad \gamma = |T|. \quad (1.8)$$

Note that these equations are coupled to the external variables through the jump  $[v]$  in  $v$ , and through the continuity of  $\sigma$ . This system of equations may be regarded as a jump condition for stationary shocks, analogous to the usual Rankine-Hugoniot condition for shocks. However, the Rankine-Hugoniot condition is a system of algebraic equations, in contrast to (1.8), which is a system of differential equations. The jump condition (1.8) effectively widens the class of weak solutions of equations (1.1) beyond the class of solutions whose jump discontinuities satisfy the usual Rankine-Hugoniot conditions. It is within this wider class that we shall seek solutions of Riemann problems. Note that system (1.8) is not scale invariant, because of the right hand side. Correspondingly, solutions of Riemann problems also fail to be scale invariant.

In Section 2, we extend results of Garaizar [1] concerning the Riemann problem, to the case in which a shear band forms. In Section 3, we describe numerical results in a test case that shows how the computations agree with the theoretical predictions.

## 2. Analytic Solution of the Riemann Problem.

In this section, we review and extend results of Garaizar [1] concerning Riemann problems for system (1.1). In Subsection 2.1, we summarize short time existence and asymptotics, which are used in Subsection 2.2 to extend the solution globally in time. The analysis applies to a simplified version of systems (1.1, 1.8), in which we linearize the yield condition (1.3), and work with perturbations of the original variables about the point at which system (1.1) loses hyperbolicity. The simplified version of equation (1.1) is

$$\begin{aligned}
\partial_t v &= \partial_x \sigma & (a) \\
\partial_t \sigma + \frac{1}{h(\gamma)} \partial_t \gamma &= c^2 \partial_x v & (b) \\
\partial_t \tau - \frac{\alpha}{h(\gamma)} \partial_t \gamma &= 0 & (c) \\
\partial_t \gamma &= \begin{cases} \partial_t(\sigma + \alpha \tau) & \text{if loading} \\ 0 & \text{otherwise,} \end{cases} & (d)
\end{aligned} \tag{2.1}$$

Here,  $\alpha = \tan \frac{\theta}{2}$ ;  $h(\gamma)$  is a positive strictly decreasing function on an interval containing the point  $\gamma = 0$ , and

$$h(0) = \alpha^2. \tag{2.2}$$

Note that in the simplified equations, we use the same symbols  $\sigma, \tau$  and  $\gamma$  to denote perturbations of the original stresses  $\sigma, \tau$  and yield stress  $\gamma$ .

## 2.1 Short time behavior.

Let  $U = (v, \sigma, \tau)^T$ , and write system (2.1a,b,c) in the loading case (in which  $\gamma = \sigma + \alpha \tau$ ) in the form

$$U_t + BU_x = 0, \tag{2.3}$$

where

$$B = -\frac{1}{h - \alpha^2 + 1} \begin{pmatrix} 0 & h - \alpha^2 + 1 & 0 \\ c^2(h - \alpha^2) & 0 & 0 \\ \alpha c^2 & 0 & 0 \end{pmatrix} \tag{2.4}$$

Characteristic speeds of (2.3) are eigenvalues of  $B$ , given by

$$\lambda_{\pm} = \pm c\sqrt{\eta}, \quad \lambda_0 = 0, \tag{2.5}$$

where  $\eta = (h - \alpha^2)/(h - \alpha^2 + 1)$ . The associated eigenvectors are (respectively)

$$r_{\pm} = (c^{-1}\sqrt{\eta}, \mp \eta, \mp \alpha(1 - \eta))^T, \quad r_0 = (0, 0, 1)^T. \tag{2.6}$$

We conclude that system (2.3) is hyperbolic if and only if  $\eta \geq 0$ . Therefore, (2.3) is hyperbolic if and only if  $\gamma \leq 0$ .

Next we describe the values of the variables within a rarefaction wave near  $\gamma = 0$ . Let  $v_0, \sigma_0, \tau_0 = -\alpha^{-1}\sigma_0, \gamma_0 = 0$  be the values of the variables at  $x = 0$  in a right moving rarefaction wave whose trailing edge has speed zero. Then (cf. [5]) we have the following expressions for the variables near  $x = 0$ , in which  $\xi = x/t$ :

$$\begin{aligned}
\gamma &= \hat{\gamma}(\xi) = -\frac{\xi^2}{c^2 h_1} + \xi^4 f_1(\xi^2), & (a) \\
v &= \hat{v}(\xi) = v_0 + v_3 \xi^3 + \xi^5 f_2(\xi^2) & (b) \\
\sigma &= \hat{\sigma}(\xi) = \sigma_0 - \sigma_4 \xi^4 + \xi^6 f_3(\xi^2), & (c) \\
\tau &= \hat{\tau}(\xi) = \tau_0 - \frac{1}{\alpha h_1 c^2} \xi^2 + \xi^4 f_4(\xi^2), & (d)
\end{aligned} \tag{2.7}$$

where

$$v_3 = \frac{2}{3\alpha^2 h_1 c^4}, \quad \sigma_4 = \frac{1}{2\alpha^2 h_1 c^4}, \quad h_1 = h'(0) < 0, \tag{2.8}$$

and the functions  $f_i$ ,  $i = 1, \dots, 4$  are real analytic near the origin if  $h$  is real analytic.

The other ingredient we require is an integrated form of the following simplified form of equations (1.8):

$$\begin{aligned}
\partial_t \sigma + \frac{1}{h(\gamma)} \partial_t \gamma &= c^2 \frac{[v]}{\delta} & (a) \\
\partial_t z - \frac{\alpha}{h(\gamma)} \partial_t \gamma &= 0 & (b) \\
\gamma &= \sigma + \alpha z. & (c)
\end{aligned} \tag{2.9}$$

**Lemma 2.1** Suppose  $h(\gamma)$  is real analytic in a neighborhood of  $\gamma = 0$ ,  $h(0) = \alpha^2 \neq 0$ , and  $h'(0) = -h_1 < 0$ . Let  $(\sigma, z, \gamma, [v])(t)$  satisfy equations (1.8), and  $(\sigma, z, \gamma)(0) = (\sigma_0, -\sigma_0/\alpha, 0)$  for some  $\sigma_0$ . Then there is a function  $b = b(\gamma)$  that is real analytic in a neighborhood of  $\gamma = 0$  such that

$$b(0) = \sigma_0, \quad b'(0) = 0, \quad b''(0) = -\alpha^2/h_1, \tag{2.10}$$

and with the property that

$$\sigma(t) = b\left(\frac{c^2 h_1}{\delta} \int_0^t [v](\theta) d\theta\right). \tag{2.11}$$

Now we are ready to consider the central problem. Consider the Riemann problem with initial data

$$U(x, 0) = \begin{cases} U_L & \text{if } x < 0 \\ U_R & \text{if } x > 0 \end{cases} \tag{2.12}$$

for which there is no centered solution of system (2.1) involving only shocks and rarefactions. Specifically, consider a combination of left moving shocks and rarefactions such that the value of  $U$  to the left of this combination is  $U_L$  and the trailing edge of the rarefaction has zero speed

(i.e.,  $\gamma = 0$ ). Similarly, consider a combination of right-moving waves, with  $U_R$  on the right, and zero speed on the left. Let  $(\sigma_l^0, v_l^0)$ ,  $(\sigma_r^0, v_r^0)$  denote the values of  $(\sigma, v)$  on the right and left (respectively) of the left and right moving wave groups (see Figure 2). In the so-called *symmetric case*, in which  $\sigma_l^0 = \sigma_r^0$ , there is no classical solution of the Riemann problem if  $v_r^0 > v_l^0$ . (That is, there is no solution involving only shocks and rarefactions.) This is the situation that we treat.

Since there is no classical solution, we explore the possibility of solving the initial value problem by including a shear band. In such a solution,  $v$  (in addition to  $\tau$  and  $\gamma$ ) can experience a jump across the  $t$ -axis. In contrast with the classical solution, the solution with a shear band is non-constant on the  $t$  axis because of equations (1.8), and the overall solution does not enjoy the property of scale invariance. To start with, consider the equation (1.8) on the shear band, which is located on the  $t$ -axis. To understand the short-time behavior, we rescale time  $t$  by a small constant:  $t' = t/\epsilon$ . Then  $\partial_t = \frac{1}{\epsilon} \partial_{t'}$ , so that the equations (1.8) are unchanged apart from an  $\epsilon$  multiplying the right hand side. Then if  $\epsilon$  is small compared with  $\delta$ , we effectively have scale invariant equations; viz.,

$$\begin{aligned} \partial_{t'} \sigma + \frac{1}{h(\gamma)} \partial_{t'} \gamma &= 0 \\ \partial_{t'} \tau - \frac{\alpha}{h(\gamma)} \partial_{t'} \gamma &= 0, \quad \gamma = \sigma + \alpha \tau. \end{aligned} \tag{2.13}$$

It follows that  $\sigma, \tau$  and  $\gamma$  are constant in time to leading order in  $\epsilon$ . Thus for small time, specifically, for times that are small compared to  $\delta$ , the solution is approximately scale invariant. In this solution, there is a rarefaction wave on the left and right extending up to the  $t$ -axis. The *blown-up* solution is shown in Figure 2.

We modify this picture for larger times as follows. The solution is no longer scale invariant, and the material unloads away from the shear band. We therefore postulate an unloading or relief wave propagating away from the shear band and interacting with the rarefaction. The conjectured structure is shown in Figure 1.

In Figure 1, the solution is scale invariant outside the region bounded by the relief waves, and agrees with the blown-up solution there. We therefore have a pair of coupled quarter-plane problems to solve in the regions  $E_l, E_r$  of Figure 5, with the boundaries in the  $(x, t)$  plane being the two relief waves and the  $t$ -axis. Note that regions  $E_l, E_r$  have  $t < T$ , where  $T$  is chosen so that the relief wave has not completely penetrated the rarefaction by time  $T$ .

Finding the solution in regions  $E_l, E_r$  reduces to solving a Goursat-type free boundary problem for the wave equation

$$\begin{aligned} \partial_t v &= \partial_x \sigma & (a) \\ \partial_t \sigma &= c^2 \partial_x v & (b) \end{aligned} \tag{2.14}$$



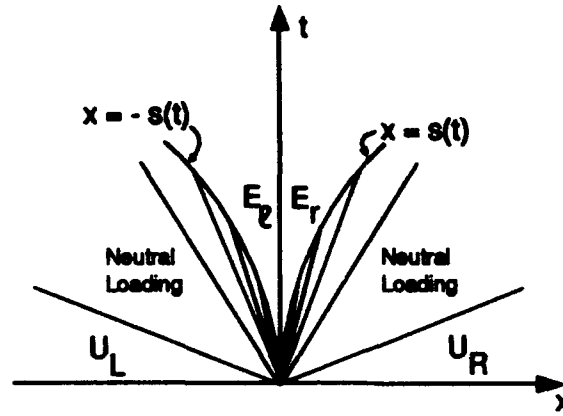


Figure 1. Solution of the Riemann Problem in the Symmetric Case.

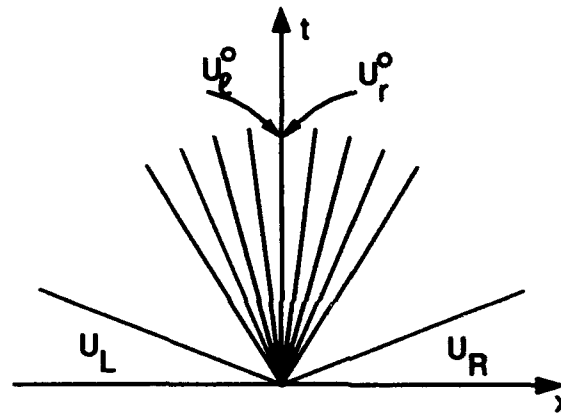


Figure 2. Blown-up Solution of the Riemann Problem.

in the planar domain

$$\{(x, t) : t > 0, 0 < x < s(t)\}. \quad (2.15)$$

At the free boundary, (2.14) is subject to two boundary conditions of Dirichlet type, corresponding to the rarefaction wave:

$$\begin{aligned} v(s(t), t) &= \hat{v}(s(t)/t) & (a) \\ \sigma(s(t), t) &= \hat{\sigma}(s(t)/t). & (b) \end{aligned} \quad (2.16)$$

At  $\{x = 0\}$ , system (2.14) is subject to the nonlinear, integral condition

$$\sigma(0, t) = b \left( \frac{2c^2 h_1}{\delta} \int_0^t v(0, t') dt' \right), \quad (2.17)$$

derived from Lemma 2.1. (Note that by adding a constant to  $v$ , we may take  $[v] = 2v$  without loss of generality.) Substituting d'Alembert's solution, written in the form

$$\begin{pmatrix} v \\ \sigma \end{pmatrix}(x, t) = F(ct + x) \begin{pmatrix} 1 \\ c \end{pmatrix} + G(ct - x) \begin{pmatrix} 1 \\ -c \end{pmatrix}, \quad 0 < x < s(t) \quad (2.18)$$

into the boundary conditions, we obtain the following three equations for the three unknown functions  $F, G, s$ :

$$\begin{aligned} F(ct + s(t)) + G(ct - s(t)) &= \hat{v}(s(t)/t) & (a) \\ F(ct + s(t)) - G(ct - s(t)) &= c^{-1} \hat{\sigma}(s(t)/t) & (b) \\ F(ct) - G(ct) &= c^{-1} b \left( \frac{2c^2 h_1}{\delta} \int_0^t (F(c\theta) + G(c\theta)) d\theta \right) & (c) \end{aligned} \quad (2.19)$$

We can derive short time asymptotic solutions of these equations, with the following result, in terms of the physical variables ([5]):

**Proposition 2.1** *For  $0 < x < s(t)$ , and near  $t = 0$ , the functions have the following expansions, uniform in  $x$ .*

$$\begin{aligned} v(x, t) &= v_0 + 2cAt^{3/2} + O(t^{5/2}) & \sigma(x, t) &= \sigma_0 + 3c^{1/2} Axt^{1/2} + Bt^2 + O(t^{5/2}) \\ \tau(x, t) &= \tau_0 - \frac{s_1^{4/3}}{h_1 \alpha} x^{2/3} + O(t^2) & \gamma(x, t) &= -\frac{s_1^{4/3}}{h_1} x^{2/3} + O(t^2), \end{aligned} \quad (2.20)$$

where

$$A = \frac{1}{3} \left( \frac{4}{3} \right)^{3/4} \alpha h_1^{1/2} c^{1/2} \left( \frac{v_0}{\delta} \right)^{3/2} > 0 \quad B = -s_1^4 \left( \sigma_4 + \frac{3}{2} v_3 \right) = -2\alpha^2 h_1 c^4 \left( \frac{v_0}{\delta} \right)^2 < 0. \quad (2.21)$$

In reference [6], we prove that there is a solution for short time that agrees with the asymptotic form we have found. The proof is based on a form of the implicit function theorem that requires the functions to be real analytic.

## 2.2 Long time behavior.

Once the local existence (for  $t < t_0$ ) of solutions has been established, we can extend these solutions to all values of  $t$  using an iterative method. That is, we can show that if solutions exist for  $0 < t \leq t'$  then these solutions can be extended to a larger interval  $0 < t \leq t' + \epsilon$  where  $\epsilon$  depends only on  $t_0$ .

Let us define a new function  $r(t)$  to replace the unknown  $s(t)$ . This function is defined implicitly by

$$r(ct + s(t)) = \frac{s(t)}{t}. \quad (2.22)$$

Then equations (2.19) become:

$$\begin{aligned} (a) \quad F + G \circ \phi &= \hat{v}(r(t)) \\ (b) \quad F - G \circ \phi &= c^{-1} \hat{\sigma}(r(t)) \\ (c) \quad F - G &= c^{-1} b \left( \frac{2ch_1}{\delta} \int_0^{t/c} (F + G) dt' \right), \end{aligned} \quad (2.23)$$

where  $\phi(t) = (c - r(t))t/(c + r(t))$ . We rearrange equations (2.23) as follows:

$$\begin{aligned} (a) \quad 2F(t) &= (\hat{v} + c^{-1}\hat{\sigma})(r(t)) \\ (b) \quad 2G(\frac{c-r(t)}{c+r(t)}t) &= (\hat{v} - c^{-1}\hat{\sigma})(r(t)) \\ (c) \quad G(t) &= F(t) - c^{-1}b(\frac{2ch_1}{\delta} \int_0^{t/c} (F + G)dt') \end{aligned} \quad (2.24)$$

Since  $r(t) \geq 0$ , we notice that  $\frac{c-r(t)}{c+r(t)}t$ , the argument of  $G$  on the LHS of (2.24b), is strictly less than  $t_0$  for all  $t$  in  $[0, t_0]$ . Therefore we can solve (2.24b) for  $r(t)$ , for values of  $t$  slightly larger than  $t_0$ . That is, we can extend  $r(t)$  to the interval  $[0, t_0 + \epsilon]$ , where  $\epsilon$  is determined from the inequality  $t - t_0 \leq \frac{2r(t)}{c-r(t)}t_0$ , if  $r(t) < c$ . Now we can use equations (2.24a) and (2.24c) to extend  $F$  and  $G$  to the same interval.

This process can be repeated indefinitely to obtain a solution to (2.23) for all  $t > 0$ , provided that the value of  $\epsilon$  does not decrease; in particular, if  $\epsilon$  depends only on the value of  $t_0$  given by the local existence results. This will be true if, for all values of  $t$ , (i)  $r(t) < c$  and (ii)  $r(t)$  is monotone increasing. The following Proposition, proved in [2] provides these properties.

**Proposition 2.2** *If equations (2.23) admit a solution  $(F, G, r)$  on some interval  $[0, T]$ , then each of these functions is monotone increasing on this interval, and  $r(t) < c$ .*

Combining this analysis with the local existence results of [5] we have established the existence of solutions for all  $t > 0$ .

### 3. Numerical Results.

In this section we will explain the numerical approach to solving system (1.1). A similar (but simplified) approach will apply to system (2.1) with a linear yield condition. Instead of describing the numerical algorithm in detail we will address some of the numerical difficulties which arise from the properties of the system.

#### 3.1 Elasto-plastic transition.

The first difficulty in designing a numerical code is that the equations in system (1.1) change depending on whether the material deforms elastically (1.1c) or plastically (1.1b). In the numerical algorithm, we add an internal variable to the set of variables  $U = (v, \sigma, \tau, \gamma)^T$  describing the material states. This variable indicates if a given material point is undergoing an elastic or plastic deformation. The internal variable is allowed to change during the course of a time update, thus avoiding the calculation of unphysical values of stress which are beyond the yield surface ( $\sigma^2 + \tau^2 > \gamma^2$ ).

#### 3.2 Stress evolution.

When the material is deforming plastically, system (1.1), (1.5) is not in conservation form. Although the momentum equation (1.1a) is a conservation law, the equation (1.1b) describing

the stress evolution during plastic deformation is not in conservation form. In order to perform a time update at a given point, we use a numerical scheme (following the ideas of Trangenstein and Colella [7]) which combines two different algorithms: (i) a second order (Godunov) hyperbolic scheme for the momentum equation and (ii) a second order implicit ordinary differential equations integrator, for the stress equation along a particle path.

The Godunov scheme for the momentum equation is :

$$v_i^{n+1} = v_i^n + c^2 \frac{\Delta t}{\Delta x} (\sigma_{i+\frac{1}{2}}^{n+\frac{1}{2}} - \sigma_{i-\frac{1}{2}}^{n+\frac{1}{2}}). \quad (3.25)$$

Here  $\sigma_{i+\frac{1}{2}}^{n+\frac{1}{2}}$  is approximated by writing the equations of motion in quasi-linear form and tracing information along characteristics. The algorithm used on the stress equations is described as follows:

Let  $\vec{S}$  be  $\begin{pmatrix} \sigma \\ \tau \\ \gamma \end{pmatrix}$ . Then  $\begin{pmatrix} \sigma_i^{n+1} \\ \tau_i^{n+1} \\ \gamma_i^{n+1} \end{pmatrix} = \vec{S}^{n+1}$  is the solution at  $t = t_{n+1}$  after the numerical integration of

$$\vec{S}_t = \frac{v_{i+\frac{1}{2}}^{n+\frac{1}{2}} - v_{i-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} \mathbf{G}(\vec{S}), \quad (3.26)$$

with initial data  $\vec{S}(t_n) = \begin{pmatrix} \sigma_i^n \\ \tau_i^n \\ \gamma_i^n \end{pmatrix}$ . Equation (3.26) and in particular  $\mathbf{G}(\vec{S})$  is derived from (1.1b,c)

and (1.5), replacing  $\partial_x v$  with  $\frac{v_{i+\frac{1}{2}}^{n+\frac{1}{2}} - v_{i-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x}$ . In both algorithms,  $\sigma_{i+\frac{1}{2}}^{n+\frac{1}{2}}$  and  $v_{i+\frac{1}{2}}^{n+\frac{1}{2}}$  are the values of  $\sigma$  and  $v$  at the cell boundaries evaluated at the intermediate time  $t = t_n + \Delta t/2$ . These quantities are computed using a characteristic tracing algorithm to achieve second order accuracy.

### 3.3 Loss of hyperbolicity.

As was noted earlier, there is a critical value of yield stress  $\gamma$ , beyond which the system is unstable. This instability relates to the loss of hyperbolicity in the full two dimensional system for signals traveling in certain directions. Physically, this loss of hyperbolicity is associated with the formation of shear bands.

In our numerical algorithm, after the time update, each cell is tested to see if the state on the cell is inside or outside the region of hyperbolicity. If the state is outside this region, a shear band is then created in the cell. From this instant, the band is treated as an internal boundary with its own equations (1.8) governing the evolution of the stress in the interior of the band.

The numerical algorithm first performs the update of the states on all the cells as if shear bands are not present. This includes the computation of stress and velocity at the cell boundaries. Next it corrects the states on cells where shear bands are present. This is done by first integrating

an ordinary differential equation similar to that of (3.26)

$$\frac{d}{dt} \vec{S}_b = \frac{v_b^R - v_b^L}{\delta} \mathbf{G}(\vec{S}_b), \text{ with initial data } \vec{S}_b^n = \begin{pmatrix} \sigma_b^n \\ \tau_b^n \\ \gamma_b^n \end{pmatrix} = \vec{S}_b(t_n) \quad (3.27)$$

where  $\delta$  is the physical width of the band,  $\vec{S}_b$  is the stress vector inside the shear band, and  $v_b^R$  and  $v_b^L$  are the velocities on the right and left sides of the band at time  $t = t_n$ .

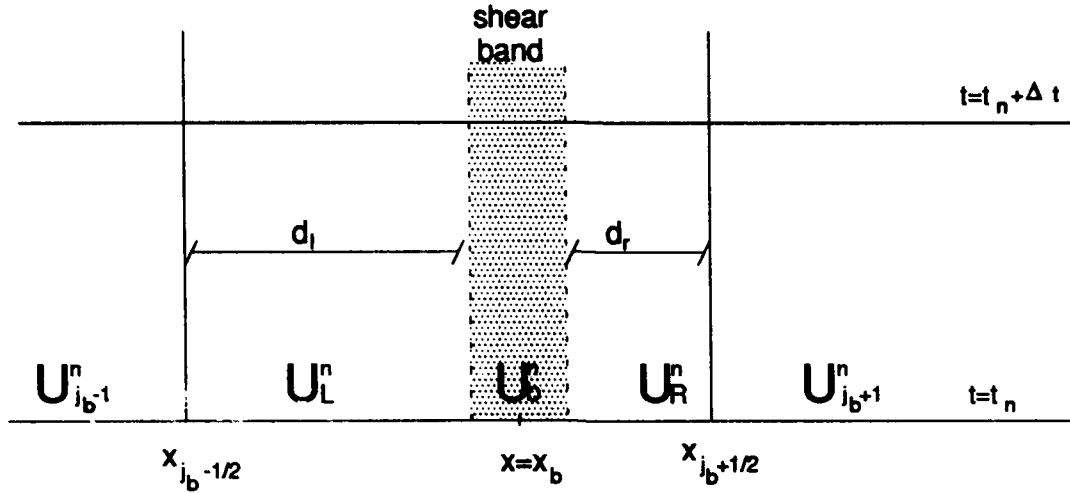


Figure 3: Cell with a shear band.

As a final step, we update the states  $U_L$  and  $U_R$  near the band; this is, on the subcells created by the formation of a shear band inside a cell (see fig 3).

The states at the band are used to compute the fluxes  $\begin{pmatrix} \sigma \\ v \end{pmatrix}$  at the fictitious boundary (i.e., shear band). The states  $U_L$  and  $U_R$  are assigned to cells of smaller size ( $d_l$  and  $d_r$ ) than a regular cell. The possible violation of the CFL condition is solved by redistributing the “numerical mass” into the nearby cells. The result of this mass redistribution is stored as a modification on the fluxes at the cell boundaries next to the shear band.

### 3.4 Numerical example

In our numerical example, we use a modification of the adaptive mesh refinement algorithm as in [8]. The computations for the following example are performed with three levels of mesh refinement, each a factor of three finer than the previous. This algorithm uses the flux information in order to assure that, during the mesh refinement process, the quantities that should be conserved are actually conserved. Thus the emphasis on expressing the mass redistribution in terms of the fluxes.

In figures 4-5 we show the profiles of the solution for a numerical example. We study an initial value problem in the interval  $0 < x < 1$ , with data

$$v = -70, \sigma = 0.6, \tau = 0, \gamma = 0.75 \text{ for } x < 0.5$$

$$v = 70, \sigma = 0.6, \tau = 0, \gamma = 0.75 \text{ for } x > 0.5$$

and with material parameters:

$$\alpha = \pi/6, c = 10^2 \text{ and } h(\gamma) = 1.7(1 - \gamma).$$

This problem does not admit a selfsimilar solution and a shear band is expected to form at  $x = 0.5$  as a result of the strong loading from both sides of the initial discontinuity.

Figure 4 shows the solution at a time soon after the formation of the shear band and Figure 5 shows the solution at a later time. The plots correspond to the stress functions  $\sigma$  and  $\gamma$  ( $\sigma \leq \gamma$  always). We observe the expected behavior of the stress at the shear band, as predicted by the analysis above (also see [5] and [2]). A precursor elastic wave travels ahead of a loading rarefaction wave which is followed by an unloading relief front. Note the coarsening of the grid in regions of neutral loading (behind the elastic precursor wave) and in regions of elastic unloading (behind the relief front).

The location of the shear band is readily identified by the dip in the values of  $\sigma$ . The nonlinearity of the yield condition ensures that the stress at the shear band will converge to a rest point under continuous loading.

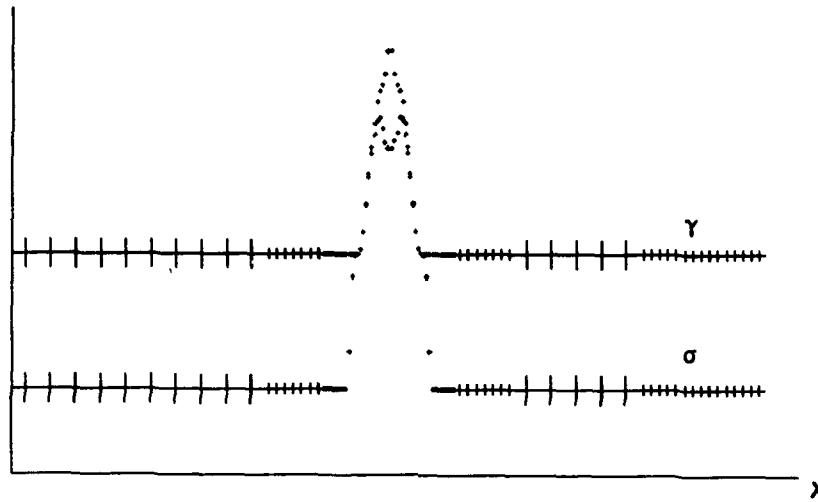


Figure 4:  $\sigma$  and  $\gamma$  at small time.

**4. Conclusions.** In this paper, we have summarized results concerning solutions of the Riemann problem for model equations describing the deformation of granular materials. The model allows for deformations with shear bands. The analytic results on existence of a solution depend upon a simplification of the equations, specifically a linearization of the yield condition about the value of stress at which a shear band forms. We outline the main features of a numerical method that is based on a higher order Godunov method, and which includes front tracking and adaptive

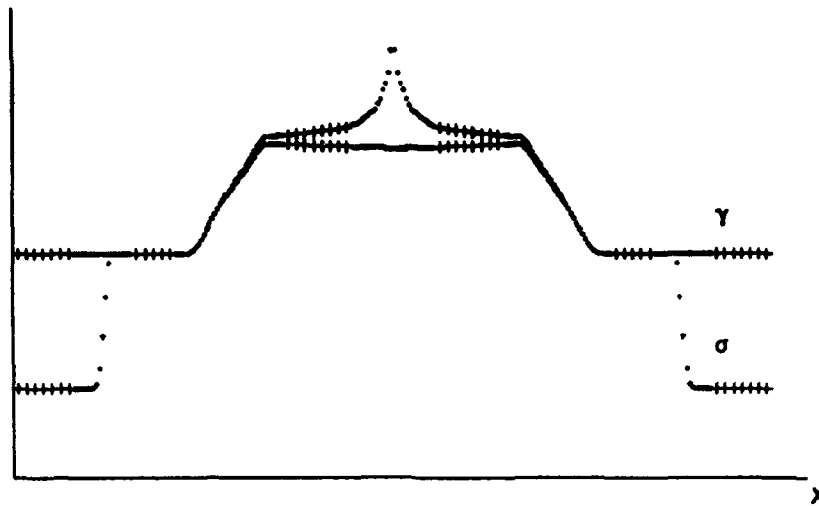


Figure 5:  $\sigma$  and  $\gamma$  at large time.

mesh refinement. This combination of techniques gives sharp resolution of the shear band, and accommodates large elastic wave speeds. The way adaptive mesh refinement coarsens and refines the grid in various regions of the material can be observed in Figures 4 and 5. The numerical algorithm performs well in capturing the features of the solution predicted by the theory.

Many of the features of the algorithm can be extended to two space dimensions, although this is somewhat complicated. The main difficulty in designing a code in two dimensions is that of capturing the growth of the shear band as the tip of the shear band propagates across the material. The algorithm in two dimensions is at an advanced stage of development.

## References

- [1] X. Garaizar, Numerical computations for antiplane shear in a granular flow model. *Quart. Appl. Math.*, to appear.
- [2] X. Garaizar and D.G. Schaeffer, Numerical Computations for shear bands in an antiplane shear model. *J. Mech. Phys. Solids*, to appear.
- [3] P.D. Lax, Hyperbolic systems of conservation laws II. *Comm. Pure Appl. Math.* **10** (1957), 537-566.
- [4] D.G. Schaeffer, A mathematical model for localization in granular flow. *Proc. Roy. Soc. Lond. A* **436** (1992), 217-250.
- [5] M. Shearer and D.G. Schaeffer, Unloading near a shear band in granular material. *Quart. Appl. Math.*, to appear.

- [6] D.G. Schaeffer and M. Shearer, Unloading near a shear band: a free boundary problem for the wave equation. *Comm. P.D.E.*, **18** (1993), 1271-1298.
- [7] J.A. Trangenstein and P. Colella, A higher-order Godunov method for modeling finite deformation in elastic-plastic solids. *Comm. Pure Appl. Math.*, **44** (1991), 41-100.
- [8] J.A. Trangenstein, Adaptive mesh refinement for wave propagation in nonlinear solids. Submitted to *SIAM J. Sci. Stat. Comput.*



# TOWARD A EUCLIDEAN ALGORITHM FOR COMPOSITION OF RATIONAL FUNCTIONS

Moss Sweedler  
ACSyAM, MSI  
Cornell University  
Ithaca NY 14853

**ABSTRACT.** Where the usual Euclidean algorithm finds a greatest common divisor, we describe a compositional Euclidean algorithm which would give a largest functional composition factor. And where the usual Euclidean algorithm is based on repeated division with remainder, the compositional Euclidean algorithm is based upon repeated compositional division with remainder. The challenge is that while we propose an apparently reasonable notion of compositional division with remainder, we do not know if it exists. However, it is easy to give a number of consequences, assuming that it does exist.

**INTRODUCTION.** The Euclidean Algorithm is a fundamental algorithm for manipulating integers and polynomials in one variable. For both integers and polynomials, the Euclidean algorithm comes down to suitably repeated division with remainder. In both settings the Euclidean algorithm produces a greatest common divisor (GCD) of two initial elements. In ideal theoretic terms, the Euclidean algorithm produces an element which generates the same ideal as the ideal generated by the two initial elements. For an ideal with a given finite number of generators, apply the Euclidean algorithm repeatedly to produce a single element which generates the same ideal.

The rational function field in one variable - denoted  $k(X)$  - consists of quotients of polynomials in the variable  $X$  with the usual arithmetic identities among fractions. The polynomials have coefficients in a field  $k$  which we refer to as the *base field*.  $k(X)$  has many proper subfields containing  $k$ , for example, consider all fractions where the numerator and denominator are polynomials with only even degree terms. This subfield is *generated* as a subfield, not as an ideal, by  $X^2$  and is naturally denoted  $k(X^2)$ . Luroth's theorem [1, p.522] says that if  $B$  is a subfield of  $k(X)$  containing  $k$  then  $B$  is generated over  $k$  by a single element; i.e.  $B = k(b)$  for some suitable element  $b$ . This is the starting point of the present paper. It is natural to consider the possibility of an algorithm *A in the spirit of the Euclidean algorithm* which does the following: given two rational functions  $b$  and  $c$ , the algorithm produces  $d$  where  $k(b,c) = k(d)$ . We wish to emphasize that we are not trying to produce the first algorithmic solution to this problem, rather we are trying to stimulate new ways of working with rational functions. For *A* to be in the spirit of the Euclidean algorithm it must be based on an algorithm *B* which plays the role that division with remainder plays in the Euclidean algorithm.

Before presenting greater detail about the possible form of *A* and *B*, here is another view of the problem. In this view functional composition is the analog to product of integers or polynomials. Of course product is commutative and composition is not. That is why in a composite such as  $f(g(X))$  we refer to  $f$  as the outer compositional element and  $g$  as the inner compositional element (ICE). The reason to switch from

\* Supported by the U.S. Army Research Office.

product to composite results from the difference between ideal generation and subfield generation. The subfield generated by  $d$  consists of rational functions which can be expressed  $f(d)/g(d)$  where  $f$  and  $g$  are polynomials in one variable.  $f/g$  is a rational function  $r$  in one variable. Hence the subfield consists of all rational functions which have  $d$  as ICE in some rational function decomposition. This is the analog to an ideal generated by a single element. The subfield generated by  $b$  and  $c$  consists of rational functions which can be expressed  $F(b,c)/G(b,c)$  where  $F$  and  $G$  are polynomials in two variables.  $F/G$  is a rational function  $R$  in two variables. If  $k(b,c) = k(d)$  it follows that there are rational functions  $r_1$  and  $r_2$  of one variable and a rational function  $R$  of two variables where:

$$1 \quad b = r_1(d) \quad c = r_2(d) \quad d = R(b,c)$$

The first two equations show that  $d$  ICE's both  $b$  and  $c$ , meaning both  $b$  and  $c$  have  $d$  as an ICE. The third equation implies that any ICE of  $b$  and  $c$  also ICE's  $d$ . I.e. if:

$$2 \quad b = s_1(e) \quad \text{and} \quad c = s_2(e) \quad \text{then} \quad d = R(s_1(e), s_2(e)) = R(s_1, s_2)(e)$$

Thus  $d$  is a largest ICE (LICE) of  $b$  and  $c$  where LICE has roughly the defining property with respect to composition as GCD has with respect to product. In fact, (1) consists of the main three equations for GCD with composition replacing product.

**SECTION ONE.** Now let us work toward the possible form of  $B$ , the replacement for division with remainder. In both integer and polynomial division the remainder is smaller than the divisor. With integers *smaller* is with respect to magnitude. With polynomials *smaller* is with respect to degree. In order to give a compositional analogue to division with remainder, we require a suitable notion of *smaller*. Suppose  $\alpha$  is a non-zero integer. The magnitude of  $\alpha$  equals the cardinality of the set: the integers modulo the ideal generated by  $\alpha$ . If  $\alpha(X)$  is a non-zero polynomial, the degree of  $\alpha$  equals the dimension as a vector space over  $k$  of  $k[X]$  modulo the ideal generated by  $\alpha(X)$ . In both cases the *size* of  $\alpha$  is determined by a relative measure of the ideal generated by  $\alpha$  to the entire ring. We use such a measure for rational functions. If  $\alpha(X)$  is a non-constant rational function, the size of  $\alpha$  is defined to be the dimension of  $k(X)$  as a vector space over the subfield  $k(\alpha(X))$ ; i.e.  $[k(X):k(\alpha(X))]$ . This integer is called the *ice degree* of  $\alpha(X)$ . By [1, p.520, thm 8.38] if  $\alpha(X)$  is written in the form  $f(X)/g(X)$  with  $f$  and  $g$  relatively prime polynomials, the maximum of the usual polynomial degrees of  $f$  and  $g$  equals  $[k(X):k(\alpha(X))]$ . Hence, the ice degree is also determined directly from  $\alpha(X)$  and is not only a relative concept.

If  $\alpha$  is zero the cardinality of the integers modulo the ideal generated by  $\alpha$  is *infinity*. Same for dimension over  $k$  of  $k[X]$  modulo the ideal generated by  $\alpha(X)$  when  $\alpha(X)$  is the zero polynomial. The magnitude or degree is in some sense opposite to the size of the quotient in these exceptional cases. Ice degree exhibits the same anomaly. When  $\alpha(X)$  is a constant rational function,  $k(\alpha(X)) = k$  so that  $[k(X):k(\alpha(X))]$  is *infinity*. The ice degree of a constant rational function is zero by convention, which accords with the maximum of the usual polynomial degrees of the numerator and denominator.

The defining formula for division with remainder, dividing  $\alpha$  by  $\beta$ , is:  $\alpha = \Gamma * \beta + \rho$ , where  $\Gamma$  is the quotient and  $\rho$  the remainder. A naive compositional analog to division with remainder is given by:  $\alpha = \Omega(\beta) + \rho$ . We have replaced the product "\*" by composition, but the sum "+" and general shape of the equation is unchanged. Let us

take a slightly more sophisticated approach. Division with remainder is effective in the realm of ideal theory. The reason is because an equation of the form:  $\alpha = \Gamma * \beta + \rho$  is universally, arithmetically reversible to give  $\rho = \alpha - \Gamma * \alpha$ . Consequently the ideal generated by  $\alpha$  and  $\beta$  equals the ideal generated by  $\beta$  and  $\rho$ . This is the key property upon which the Euclidean algorithm is based and which we must preserve. In other words, if *compositional division* of  $\alpha$  by  $\beta$  is to produce an *ice remainder*  $\rho$ , the process ought to be universally, compositionally reversible so that  $k(\alpha, \beta) = k(\beta, \rho)$ . Once we achieve this, the compositional Euclidean algorithm follows. Let us play with this condition:  $k(\alpha, \beta) = k(\beta, \rho)$ . Collecting a common  $k(\beta)$  we are asking:  $k(\beta)(\alpha) = k(\beta)(\rho)$ . Fractional linear transformations are automorphisms of function fields. It is natural to seek  $a, b, c, d$  in  $k(\beta)$  where:

- 3  $ad - bc$  is non-zero and the ice degree of  $(a\alpha + b) / (c\alpha + d)$  is less than the ice degree of  $\beta$ . Let  $\rho$  denote  $(a\alpha + b) / (c\alpha + d)$ .

The non-zero (determinant) condition on  $ad - bc$  insures that

- 4  $(e\rho + f) / (g\rho + h) = \alpha$  for suitable  $e, f, g, h$  in  $k(\beta)$ ; hence  $k(\beta)(\alpha) = k(\beta)(\rho)$

**DEFINITION** Given rational functions  $\alpha$  and  $\beta$  where  $\beta$  has ice degree at least one, the *compositional division* of  $\alpha$  by  $\beta$  consists of  $a, b, c, d$  in  $k(\beta)$  satisfying (3). For such a compositional division,  $\rho$  is the *ice remainder*. The fractional linear transformation and rational function:  $(aX + b)/(cX + d)$  is the *outer compositional quotient* of  $\alpha$  by  $\beta$ .

Here are three questions concerning compositional division, the first of which is: is compositional division always possible? For example, suppose  $\beta$  has ice degree at least one and  $\alpha$  does not lie in  $k(\beta)$ . In this case  $k(\beta)(\alpha)$  properly contains  $k(\beta)$ . By Luroth's theorem,  $k(\beta)(\alpha)$  is generated by a single element  $\rho$ . Since  $k(\rho) = k(\beta)(\alpha)$  which properly contains  $k(\beta)$ , it follows that  $\rho$  has lower ice degree than  $\beta$ . It is clear that  $\rho$  can be written as a polynomial in  $\alpha$  with coefficients from  $k(\beta)$  but can  $\rho$ , or some other element of  $k(\beta)(\alpha)$  with lower ice degree than  $\beta$ , be written as a fractional linear transformation of  $\alpha$  with coefficients from  $k(\beta)$ ? The next two questions assume that compositional division is possible. In this case: a. give a *beautiful* algorithm for compositional division, or at least for finding  $\rho$ ; b. give an efficient algorithm for compositional division, or at least for finding  $\rho$ .

We conclude by deriving consequences of compositional division, including the compositional Euclidean algorithm.

**LEMMA** Given rational functions  $\alpha$  and  $\beta$  where  $\beta$  has ice degree at least one, the *compositional division* of  $\alpha$  by  $\beta$  exists and has ice remainder in  $k$  if and only if  $\alpha$  lies in  $k(\beta)$ . In this case any element of  $k$  can be achieved as an ice remainder.

**PROOF** Say  $\alpha$  lies in  $k(\beta)$  and  $s$  is any element of  $k$ . In (3) let  $a = 1$ ,  $b = \alpha(s - 1)$ ,  $c = 0$  and  $d = \alpha$ . Then (3) is satisfied and  $\rho = s$ . Conversely suppose the compositional division exists with ice remainder  $\rho$  in  $k$ . From (4) and that  $\rho$  lies in  $k$  it follows that  $\alpha$  lies in  $k(\beta)$ . qed

Let us now assume that compositional division does in fact exist and show how to utilize it in the manner that ordinary division is utilized.

**PROPOSITION** Let  $K$  be a subfield of  $k(X)$  which properly contains  $k$ . Then  $K$  contains elements of non-zero ice degree. If  $\beta$  is an element of  $K$  with smallest non-zero ice degree then  $K = k(\beta)$ .

**PROOF** Any element of  $K$  lying outside of  $k$  has non-zero ice degree. Let  $\beta$  be any such element with minimal ice degree. Let  $\alpha$  be any element of  $K$ .  $\rho$  the ice remainder from compositional division of  $\alpha$  by  $\beta$  has smaller ice degree than  $\beta$ . By the minimality property of  $\beta$ , it follows that  $\rho$  has ice degree zero and hence lies in  $k$ . By the previous lemma it follows that  $\alpha$  lies in  $k(\beta)$ . qed

**COMPOSITIONAL EUCLIDEAN ALGORITHM** Starting with rational functions  $\alpha_0$  and  $\alpha_1$  construct the sequence  $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_n$  where  $\alpha_{i+1}$  is the ice remainder of compositional division of  $\alpha_{i-1}$  by  $\alpha_i$ . The process continues until reaching  $\alpha_n$  of ice degree zero. We allow the possibility that  $n = 1$  and no division is performed. Then  $\alpha_{n-1}$  is the ICE of  $\alpha_0$  and  $\alpha_1$ .

**PROOF** At each stage  $k(\alpha_0, \alpha_1) = k(\alpha_1, \alpha_2) = \dots = k(\alpha_{n-1}, \alpha_n) = k(\alpha_{n-1})$  where the last equality follows since  $\alpha_n$  lies in  $k$ . qed

#### **BLIOGRAPHY**

1. N.Jacobson, Basic Algebra II, 2nd ed. W.H.Freeman and Co. New York, 1989.

# Decoding Hyperbolic Cascaded Reed–Solomon Codes

Keith Saints

Center for Applied Mathematics  
Cornell University, Ithaca NY 14853  
email: keith@cam.cornell.edu

## Abstract

In this paper, we define Hyperbolic Cascaded Reed–Solomon (HCRS) codes, study their algebraic properties, and describe an algorithm for decoding them up to their full error-correcting capability. The codewords of HCRS codes are represented either as arrays or as bivariate polynomials. Our decoding algorithm is an extension of an algorithm of Sakata, and the decoding is performed by calculating a Gröbner basis for an ideal related to the error locations.

## Section 1. Linear Block Codes

We begin with a brief introduction to the theory of error-correcting codes, and in particular, linear block codes [1, 6].

**Error-Correcting Codes.** Error-correcting codes are used to protect digital information during transmission across a noisy channel. Applications include modem communications over a phone line, radio communications with satellites or spacecraft, and computer storage devices. In each case, noise is introduced to the data transmitted so that the bits received may not be the same as the bits transmitted. An error-correcting code adds redundancy to the information to create codewords which may be reconstructed by the receiver even if some of the bits are in error.

**Linear Block Codes.** Let  $q$  be a power of a prime number. We use the elements of  $\mathbb{F}_q$ , the finite field with  $q$  elements, as the symbols of an alphabet used to compose codewords. Usually,  $q$  is a power of 2 so that symbols can be expressed as strings of bits. The encoder takes a word of length  $k$  and encodes it as a *codeword* of length  $n > k$ . The encoder thus is a one-to-one map  $\mathbb{F}_q^k \rightarrow \mathbb{F}_q^n$ . The *code*  $C$  is the set of all codewords; that is, the image of  $\mathbb{F}_q^k$  under the encoding map. We require the encoder to be a linear map of vector spaces over  $\mathbb{F}_q$ , so that  $C$  is a *linear block code*. In this case,  $C$  is a  $k$ -dimensional subspace of  $\mathbb{F}_q^n$ .

**Hamming Distance.** The *Hamming weight*  $\|w\|_H$  of a word  $w \in \mathbb{F}_q^n$  is the number of nonzero entries of  $w$ . The *Hamming distance*  $d_H(v, w)$  between two elements  $v, w \in \mathbb{F}_q^n$  is defined to be  $\|v - w\|_H$ ; that is, the number of entries in which they disagree. The *minimum distance* of a code  $C$  is the minimum Hamming distance between any two distinct codewords of  $C$ .

**Correction of Errors.** Suppose  $C$  is a linear block code with minimum distance at least  $2t + 1$ . A codeword  $c \in C$  is sent through the channel and some of its entries are altered. Thus the receiver receives a word  $w \in \mathbb{F}_q^n$  which is the sum  $c + e$  of the codeword and an error word  $e \in \mathbb{F}_q^n$ . The number of places in which  $c$  has been corrupted

---

This work was supported by the U.S. Army Research Office through the Army Center of Excellence for Symbolic Methods in Algorithmic Mathematics (ACSyAM), Mathematical Sciences Institute of Cornell University. Contract DAAL03-92-G-0126.

is expressed as the Hamming distance  $d_H(w, c) = \|e\|_H$ . Assuming that no more than  $t$  errors occurred, we have  $d_H(w, c) \leq t$ , and  $c$  is the unique codeword with this property since by the triangle inequality,  $d_H(w, c') \geq t + 1$  for all other codewords  $c'$ .

Thus a code  $C$  with minimum distance  $2t + 1$  is a  $t$ -error correcting code, that is, any pattern of up to  $t$  errors may be corrected. The task of the decoder is to produce the codeword nearest in Hamming distance to the received word. However, the number of codewords is exponential in the blocklength, as well as the number of correctable error patterns, so finding an efficient decoding algorithm is problematic.

**Parameters of a Linear Block Code.** An  $(n, k, d)$  linear block code  $C$  is a  $k$ -dimensional subspace of  $\mathbb{F}_q^n$  with minimum distance  $d$ . The parameter  $n$  is called the *blocklength* of the code. The parameter  $k$  is called the *dimension* of the code. The *rate* of the code is the ratio  $k/n$ , since each word of  $k$  symbols is encoded as a word of  $n$  symbols. The parameter  $d$  is the *minimum distance* of the code. We have seen that a code with minimum distance  $d$  can correct up to  $t = \lfloor \frac{d-1}{2} \rfloor$  errors. In choosing a code for transmission of data across a given channel, there must be some tradeoff between the conflicting goals of high rate and high error-correcting capability. We must also keep in mind that the code ought to have an efficient decoding algorithm.

## Section 2. Hyperbolic Cascaded Reed-Solomon Codes

Hyperbolic Cascaded Reed-Solomon (HCRS) codes have been studied in [4, 5, 7, 10]. We give an algebraic description of HCRS codes here, but they were originally introduced in [10] using the cascade code construction of Blokh and Zyablov [2]. HCRS codes are in many ways a generalization of the widely-used Reed-Solomon (RS) codes [1, 6]. One motivation for using HCRS codes is their *long blocklengths*: whereas RS codes over the alphabet  $\mathbb{F}_q$  are limited to blocklengths  $\approx q$ , HCRS codes have blocklengths  $\approx q^2$ .

**Notation.** The set of nonzero elements of  $\mathbb{F}_q$  forms a cyclic group under multiplication. Let  $n = q - 1$  be the order of the cyclic group  $\mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\}$ , and choose  $\alpha \in \mathbb{F}_q$  to be a generator of this group. Thus  $\mathbb{F}_q^* = \{1, \alpha, \alpha^2, \dots, \alpha^{n-1}\}$ . Let  $\mathbb{F}_q^{n \times n}$  be the set of  $n \times n$  arrays with entries from the field  $\mathbb{F}_q$ , and let  $\mathbb{F}_q^{n \times n}[x, y]$  denote the set of bivariate polynomials  $f \in \mathbb{F}_q[x, y]$  with  $\deg_x f < n$  and  $\deg_y f < n$ . We identify  $\mathbb{F}_q^{n \times n}$  with  $\mathbb{F}_q^{n \times n}[x, y]$  by identifying a polynomial  $a(x, y)$  with the array  $a$  of its coefficients:

$$a = \begin{pmatrix} a_{00} & \cdots & a_{0,n-1} \\ \vdots & \ddots & \vdots \\ a_{n-1,0} & \cdots & a_{n-1,n-1} \end{pmatrix}$$

$$a(x, y) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a_{ij} x^i y^j$$

**The Fourier Transform.** The *Fourier transform* is a one-to-one linear map from  $\mathbb{F}_q^{n \times n}[x, y]$  (the time domain) to  $\mathbb{F}_q^{n \times n}[X, Y]$  (the frequency domain). (Of course, these two spaces are isomorphic, but we express the polynomials in different sets of variables in

order to distinguish them.) The Fourier transform,  $a \mapsto A$ , and its inverse are given by the formulas:

$$A(X, Y) = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a(\alpha^i, \alpha^j) X^i Y^j,$$

$$a(x, y) = \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} A(\alpha^{-i}, \alpha^{-j}) x^i y^j.$$

Note that  $A_{ij} = a(\alpha^i, \alpha^j)$ , so  $A$  is a *table* of the *values* of  $a(x, y)$ . The properties of the Fourier transform are studied in detail in [1].

**2-D Cyclic Codes [7, 9].** Let  $P \subset \mathbb{Z}_+^2$  be a set of “frequencies”. A *2-D cyclic code* is defined by taking as codewords those words  $a \in \mathbb{F}_q^{n \times n}$  whose Fourier transform has zero entries in the positions specified by  $P$ .

$$C_P = \{a \in \mathbb{F}_q^{n \times n} : A_{ij} = 0 \text{ for each } (i, j) \in P\}.$$

The blocklength of this code is  $N = n^2$ , since it is a subspace of  $\mathbb{F}_q^{n \times n}$ .

**Hyperbolic Cascaded Reed–Solomon Codes.** For arbitrary choices of the constraint set  $P$ , the minimum distance of the 2-D cyclic code  $C_P$  may not be very good (and difficult to determine). We define Hyperbolic Cascaded Reed–Solomon codes which can be shown to have good distance properties. Given a parameter  $d$ , the desired minimum distance, we define the frequency set

$$P_d = \{(i, j) \in \mathbb{Z}_+^2 : (i+1)(j+1) < d, i < n, j < n\}.$$

Then we define the *Hyperbolic Cascaded Reed–Solomon* code

$$\text{HCRS}_d = C_{P_d} = \{a \in \mathbb{F}_q^{n \times n} : A_{ij} = 0 \text{ whenever } (i+1)(j+1) < d\}.$$

The parameters of the code  $\text{HCRS}_d$  are  $(N = n^2, k = n^2 - |P_d|, d)$ . The dimension is  $n^2 - |P_d|$  since each constraint imposed by an element of  $P_d$  is independent, and the minimum distance can be shown to be at least  $d$ , the designed minimum distance [7].

**Examples of HCRS codes.** These ideas are best understood by looking at an example. Consider the  $(49, 35, 7)$  code  $\text{HCRS}_7$  over  $\mathbb{F}_8$ . The frequency set is given by

$$P_7 = \{(i, j) \in \mathbb{Z}_+^2 : (i+1)(j+1) < 7\}$$

$$= \{(0, 0), (0, 1), (0, 2), (0, 3), (0, 4), (0, 5), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (3, 0), (4, 0), (5, 0)\}$$

A  $7 \times 7$  array  $(a_{ij})$  is a codeword if and only if its transform  $A_{ij}$  has the form:

$$\begin{pmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} & a_{05} & a_{06} \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{50} & a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} \\ a_{60} & a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & A_{06} \\ 0 & 0 & 0 & A_{13} & A_{14} & A_{15} & A_{16} \\ 0 & 0 & A_{22} & A_{23} & A_{24} & A_{25} & A_{26} \\ 0 & A_{31} & A_{32} & A_{33} & A_{34} & A_{35} & A_{36} \\ 0 & A_{41} & A_{42} & A_{43} & A_{44} & A_{45} & A_{46} \\ 0 & A_{51} & A_{52} & A_{53} & A_{54} & A_{55} & A_{56} \\ A_{60} & A_{61} & A_{62} & A_{63} & A_{64} & A_{65} & A_{66} \end{pmatrix}$$

codeword  $a$  (time domain)                      its transform  $A$  (frequency domain)

This diagram also suggests a method of encoding: a message consisting of 35 symbols could be entered as the 35 free entries of the transform array  $A$ , and the inverse transform could then be applied to obtain a codeword. Thus we have a linear map  $\mathbb{F}_8^{35} \rightarrow \mathbb{F}_8^{49}$  whose image is  $\text{HCRS}_7$ .

**HCRS codes in  $m$  dimensions.** Consider  $m$ -dimensional  $n \times n \times \cdots \times n$  arrays over  $\mathbb{F}_q$ . Use the  $m$ -dimensional Fourier transform to define codes with blocklength  $N = n^m$ . The one-dimensional version of a HCRS code is the well-known Reed-Solomon code. For example:

$$\begin{aligned} \text{Reed-Solomon: } \text{RS}_d &= \{a \in \mathbb{F}_q^n : A_i = 0, (i+1) < d\} \\ \text{2-D HCRS: } \text{HCRS}_d &= \{a \in \mathbb{F}_q^{n \times n} : A_{ij} = 0, (i+1)(j+1) < d\} \\ \text{3-D HCRS: } \text{HCRS}_d &= \{a \in \mathbb{F}_q^{n \times n \times n} : A_{ijk} = 0, (i+1)(j+1)(k+1) < d\} \end{aligned}$$

### Section 3. Decoding HCRS Codes

We have developed an algorithm for correcting HCRS codes up to their full error-correcting capacity. That is, our algorithm corrects any pattern of  $t$  errors for the code  $\text{HCRS}_{2t+1}$ . The algorithm is given in [7]. We sketch here some of the underlying ideas.

**The Syndrome Array.** From now on, we will be considering the code  $\text{HCRS}_{2t+1}$ . A codeword  $c \in \text{HCRS}_{2t+1}$  is sent through the channel. An error  $e \in \mathbb{F}_q^{n \times n}$  may be introduced, resulting in a received word which is the sum  $w = c + e$ . Recall that the task of the decoder is to determine the unique codeword  $c$  which differs from  $w$  in at most  $t$  entries. Apply the Fourier transform:  $W_{ij} = C_{ij} + E_{ij}$ , for all  $(i, j)$ . The transform  $E$  of  $e$  is the *syndrome array*. Note that for each  $(i, j) \in P_{2t+1}$ , an entry  $E_{ij}$  of the syndrome array is known to the decoder, since  $C_{ij} = 0$  and therefore  $W_{ij} = E_{ij}$ . Thus the syndrome array is partially known, and the decoding problem may be reformulated: the decoder must find the unique completion of the partially-known syndrome array to a full  $n \times n$  array which is the transform of an error pattern with weight  $t$  or less.



**2-D Linear Recursion Relations.** Extend the syndrome array  $E$  to an infinite array:

$$E_{ij} = e(\alpha^i, \alpha^j), \quad \text{for all } (i, j) \in \mathbb{Z}_+^2.$$

Since  $\alpha^n = 1$ , this array is doubly periodic:

$$E_{i+n, j} = E_{ij}, \quad E_{i, j+n} = E_{ij}$$

This is an example of a two-dimensional linear recursion (2-D LR) relation [8]. In general, if

$$\sum_r \sum_s f_{rs} E_{i+r, j+s} = 0, \quad \text{for all } (i, j) \in \mathbb{Z}_+^2,$$

we say that  $E$  satisfies a 2-D LR relation with coefficients  $f_{rs}$ . The characteristic polynomial of this relation is:

$$f(x, y) = \sum_r \sum_s f_{rs} x^r y^s.$$

As mentioned above, the periodicity conditions  $E_{i+n, j} = E_{ij}$  and  $E_{i, j+n} = E_{ij}$  are examples of 2-D LR relations which are satisfied by every syndrome array. These relations have  $x^n - 1$  and  $y^n - 1$  as their characteristic polynomials.

**The Error Locator Ideal.** Consider the error polynomial  $e(x, y) = \sum_i \sum_j e_{ij} x^i y^j$ . For an error pattern which is correctable, there are at most  $t$  coefficients  $e_{ij}$  which are nonzero. For each nonzero  $e_{ij}$ ,

$$\begin{aligned} (i, j) & \text{ is an error location} \\ e_{ij} & \text{ is an error value.} \end{aligned}$$

The following theorem connects the error locations with the set of 2-D LR relations valid on the syndrome array.

**Error Location Theorem (Sakata [9]).**

$$F(X, Y) \text{ is a 2-D LR relation valid on } E$$

$$\iff$$

$$F(\alpha^i, \alpha^j) = 0 \text{ for each error location } (i, j).$$

This motivates us to define the *error locator ideal*:

$$\begin{aligned} L &= \{F \in \mathbb{F}_q[X, Y] : F \text{ is a valid 2-D LR relation on } E\} \\ &= \{F \in \mathbb{F}_q[X, Y] : F \text{ vanishes at each } (\alpha^i, \alpha^j)\} \end{aligned}$$

To identify the error locations, we seek a Gröbner basis [3] which generates  $L$ :

$$L = \langle F_1(X, Y), F_2(X, Y), \dots, F_t(X, Y) \rangle.$$

(In the case of Reed-Solomon codes, we are considering an ideal in the ring of polynomials in one variable. In this case, the ideal will always be principal:  $L = \langle \Lambda(X) \rangle$ , and here  $\Lambda(X)$  is the error-locator polynomial as studied in the theory of Reed-Solomon codes [1, 6].)

## Overview of Decoding Algorithm [7].

1. Find 2-D LR relations satisfied by the syndrome array. (Form a Gröbner basis for the error locator ideal  $L$ .) There are two basic operations:
  - a. Testing a 2-D LR relation to see if it correctly predicts the value of a known syndrome.
  - b. Predicting the value of an unknown syndrome using a 2-D LR relation.
2. Find the common zeros of the polynomials in  $L$ . These are the common zeros of the polynomials in the Gröbner basis. Each zero  $(\alpha^i, \alpha^j)$  identifies an error location  $(i, j)$ .
3. Interpolate to find the error values. Subtract the error from the received word to obtain the codeword.

**Example.** We again consider the  $(49, 35, 7)$  code over  $F_8$ . This code is capable of correcting any pattern of three errors or less. Choose  $\alpha$  (the primitive 7<sup>th</sup> root of unity) to be a solution of the equation  $\alpha^3 + \alpha + 1 = 0$ . We receive the following word, and calculate the corresponding syndrome array (the symbol '\*' denotes an unknown entry):

$$\begin{array}{c}
 \begin{pmatrix} \alpha^3 & \alpha^5 & \alpha^2 & \alpha^3 & \alpha^2 & \alpha^5 & 0 \\ \alpha^3 & \alpha^6 & \alpha & \alpha^5 & \alpha^4 & \alpha^6 & \alpha^4 \\ 0 & 1 & \alpha^3 & \alpha^5 & \alpha^2 & 1 & 0 \\ \alpha^4 & 0 & \alpha^5 & \alpha^4 & \alpha^3 & 1 & \alpha^2 \\ \alpha^5 & 1 & \alpha^3 & \alpha^6 & \alpha^2 & \alpha^4 & \alpha \\ \alpha^2 & \alpha^6 & 0 & 1 & \alpha^4 & \alpha^6 & \alpha \\ \alpha^5 & \alpha^2 & \alpha & \alpha^5 & 1 & \alpha^2 & \alpha^3 \end{pmatrix} \\
 \text{Received Word}
 \end{array}
 \longrightarrow
 \begin{array}{c}
 \begin{pmatrix} \alpha^4 & \alpha^3 & \alpha^3 & \alpha^6 & 0 & \alpha^4 & * \\ \alpha^3 & \alpha^5 & \alpha^3 & * & * & * & * \\ \alpha^3 & 1 & * & * & * & * & * \\ \alpha^5 & * & * & * & * & * & * \\ 0 & * & * & * & * & * & * \\ 0 & * & * & * & * & * & * \\ * & * & * & * & * & * & * \end{pmatrix} \\
 \text{Syndrome}
 \end{array}$$

The error locator ideal is found to be

$$L = \langle f_1, f_2, f_3 \rangle,$$

where

$$\begin{aligned}
 f_1 &= Y^2 + \alpha^2 Y + \alpha^5 \\
 f_2 &= XY + \alpha X + \alpha^5 Y + \alpha^6 \\
 f_3 &= X^2 + X + \alpha^6 Y + \alpha^5
 \end{aligned}$$

The corresponding relations which are satisfied by the syndrome array  $E$  are:

$$\begin{aligned}
 E_{i,j+2} + \alpha^2 E_{i,j+1} + \alpha^5 E_{i,j} &= 0 \\
 E_{i+1,j+1} + \alpha E_{i+1,j} + \alpha^5 E_{i,j+1} + \alpha^6 E_{i,j} &= 0 \\
 E_{i+2,j} + E_{i+1,j} + \alpha^6 E_{i,j+1} + \alpha^5 E_{i,j} &= 0
 \end{aligned}$$

We solve for the three common roots of the polynomials  $f_1, f_2, f_3$ :

$$(\alpha, \alpha), \quad (\alpha^3, \alpha), \quad (\alpha^5, \alpha^4).$$

This indicates that the error locations are (1,1), (3,1) and (5,4). Thus  $e(x,y) = e_{11}xy + e_{31}x^3y + e_{54}x^5y^4$ . We recalculate the syndromes in terms of this expression to solve for the error values:

$$e_{11} = \alpha^4, \quad e_{31} = 1, \quad e_{54} = 1.$$

These three values are subtracted from the corresponding entries of the received word to obtain the codeword:

$$c = \begin{pmatrix} \alpha^3 & \alpha^5 & \alpha^2 & \alpha^3 & \alpha^2 & \alpha^5 & 0 \\ \alpha^3 & \boxed{\alpha^3} & \alpha & \alpha^5 & \alpha^4 & \alpha^6 & \alpha^4 \\ 0 & 1 & \alpha^3 & \alpha^5 & \alpha^2 & 1 & 0 \\ \alpha^4 & \boxed{1} & \alpha^5 & \alpha^4 & \alpha^3 & 1 & \alpha^2 \\ \alpha^5 & 1 & \alpha^3 & \alpha^6 & \alpha^2 & \alpha^4 & \alpha \\ \alpha^2 & \alpha^6 & 0 & 1 & \boxed{\alpha^5} & \alpha^6 & \alpha \\ \alpha^5 & \alpha^2 & \alpha & \alpha^5 & 1 & \alpha^2 & \alpha^3 \end{pmatrix}$$

## References

- [1] R.E. Blahut. *Theory and Practice of Error Control Codes*. Addison-Wesley Publishing Company, Reading MA, 1983.
- [2] E. L. Blokh and V. V. Zyablov, "Coding of generalized cascade codes," *Probl. Info. Trans.*, vol. 10, pp. 45-50, 1974.
- [3] B. Buchberger, "Gröbner bases: An algorithmic method in polynomial ideal theory," in *Multidimensional Systems Theory: Progress, Directions and Open Problems in Multidimensional Systems* (N. K. Bose, ed.). Dordrecht, Holland: D. Reidel, 1985.
- [4] R. Krishnamoorthy and C. Heegard, "Structure and decoding of Reed-Solomon based cascade codes," in *Proc. 25th Ann. Conf. Inform. Sci. Syst.*, pp. 29-33, 1991.
- [5] R. Krishnamoorthy, *Algorithms for Capacity Computations and Algebraic Cascade Coding with Applications to Data Storage*. PhD thesis, Cornell University, Aug. 1991.
- [6] F.J. MacWilliams and N.J.A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam, 1977.
- [7] K. Saints and C. Heegard, "On Hyperbolic Cascaded Reed-Solomon codes," *The Tenth International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes (AAECC-10)*, San Juan, Puerto Rico, May 1993.
- [8] S. Sakata, "Finding a minimal set of linear recurring relations capable of generating a given finite two-dimensional array," *J. Symbolic Computation*, vol. 5, pp. 321-337, 1988.
- [9] S. Sakata, "Decoding binary 2-D cyclic codes by the 2-D Berlekamp-Massey algorithm," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1200-1203, July 1991.
- [10] J. Wu and D. J. Costello Jr., "New multi-level codes over GF(q)," *IEEE Trans. Inform. Theory*, vol. 38, pp. 933-939, Jan. 1992.

# On the Computational Content of Classical Sequent Proofs

Judith Underwood\*  
Department of Computer Science  
Cornell University  
Ithaca, NY 14853

September 13, 1993

## Abstract

This paper presents a method for extracting programs from classical sequent proofs in propositional logic. The term language is typed lambda calculus augmented by a nonlocal control operator. The control operator used in this paper is Scheme's *call/cc* (call-with-current-continuation). An advantage of using *call/cc* is that a subject reduction theorem can be proved very easily for this language.

One new feature of this work is the choice of the classical sequent calculus for propositional logic as the proof system. Terms are extracted directly from propositional sequent proofs, rather than via a translation to natural deduction proofs. A technique for representing continuations as partial proofs is developed, and reduction is described as an operation on proofs.

## 1 Introduction

In recent years, there has been a great deal of interest in the computational content of classical proofs. Since the discovery of a classical typing for the control operator  $C$  by Griffin [5], a great deal of work has been done in expressing the computations in classical proofs as programs in a lambda calculus extended by nonlocal control operators. Using the principle of the Curry-Howard isomorphism, if we augment the typed lambda calculus with an operator having the type of a classical axiom (in particular, an axiom strong enough to prove all of classical logic when added to intuitionistic logic), we will be able to extract programs from classical proofs. The question is then whether these programs actually represent a sensible computation.

Murthy answered this question in part in his thesis [6], where he showed that a classical proof of a  $\Pi_2^0$  sentence  $\phi$  can be interpreted as a program in an extended lambda calculus which meets the specification  $\phi$ . This work was done considering various fixed, normalizing evaluation strategies, however; the question of strong normalization for languages with control operators remains open. Barbanera and Berardi proved a strong normalization result for one such language in [1], with some substantial restrictions on the types allowed in the language. Specifically, their language forbids types with strict subtypes of the form  $\neg\neg P$ , and forbids  $\perp$  from appearing on the left hand side of an implication.

This paper presents a method for extracting programs from classical sequent proofs in propositional logic. The term language is typed lambda calculus augmented by a nonlocal control operator. The control operator used in this paper is Scheme's *call/cc* (call-with-current-continuation). An advantage of using *call/cc* is that a subject reduction theorem can be proved very easily for this language.

One unusual aspect of this work is the choice of the classical sequent calculus for propositional logic as the proof system. Terms are extracted directly from propositional sequent proofs, rather than via a translation to natural deduction proofs. A technique for representing continuations as partial proofs is developed, and reduction is described as an operation on proofs. Using this proof structure leads to a greater understanding of reduction, which may in turn lead to a strong normalization proof for these terms.

---

\*Supported in part by the United States Army Research Office through the Army Center of Excellence for Symbolic Methods in Algorithmic Mathematics (ACSyAM), Mathematical Sciences Institute of Cornell University. Contract DAAL 03-91-C-0027, and in part by a National Science Foundation Graduate Fellowship.

The work began as an attempt by the author to understand exactly how cut elimination as reduction corresponds to normalization of natural deduction proofs, which was inspired in part by Girard's work [4] on constructive interpretations of classical logic, and in part by considering an extension of the author's previous work on the tableau algorithm for intuitionistic propositional calculus ([7]).

The rest of the paper describes the details required for the proof. First, we define term language we use to express the computations involved. Then, an outline of the method of extraction of a term from a sequent proof is given. Next, we extend the computation system by the addition of typed constants, and we define the reduction rules for program terms built from terms extracted from proofs and these constants. (These rules are actually described in the form of a term rewriting machine.) The remainder of the paper describes the correspondence between reduction and cut elimination.

## 2 Term Language for Classical Logic

Each rule of the classical sequent calculus will correspond to the construction of a term in a term language based on typed lambda calculus plus call/cc, a nonlocal control operator. Classical typings for control operators were discovered by Griffin and developed by Murthy ([5, 6]). In particular, call/cc can be given the type  $(\neg P \rightarrow P) \rightarrow P$ , a classical tautology which is a form of Peirce's law. The results given here could also be expressed in a term language which uses the control operator  $C$  and the abort operator  $A$  with their operational semantics as described in [3], at the expense of increased complexity of the proofs. (To make this translation use the definition

$$\text{call/cc}(M) = (C\lambda k.k(Mk))$$

as in ([3]); this definition also justifies the typing of call/cc if we consider the usual typing of  $C$  as  $\neg(\neg P) \rightarrow P$ .)

The basic term language, with the exception of call/cc, is a subset of the Nuprl term language ([2]). We use logical notation for types, and for each type, we describe how to build terms which inhabit that type from inhabitants of its constituent types. Such terms are called *canonical* because they are not themselves reducible (though their subterms may be).

The types, and some of the terms, are defined as follows:

- Propositional letters  $(p, q, r \dots)$  are types, and their canonical inhabitants are variables  $(x, y, z, \dots)$  of the appropriate type.
- If  $P_1$  and  $P_2$  are types, then  $P_1 \wedge P_2$  is a type. If  $f_1$  and  $f_2$  are terms of type  $P_1$  and  $P_2$  respectively, then  $\text{pair}(f_1, f_2)$  is a canonical inhabitant of  $P_1 \wedge P_2$ .
- If  $P_1$  and  $P_2$  are types, then  $P_1 \vee P_2$  is a type. If  $f_1$  and  $f_2$  are terms of type  $P_1$  and  $P_2$  respectively, then  $\text{inl}(f_1)$  and  $\text{inr}(f_2)$  are canonical inhabitants of  $P_1 \vee P_2$ .
- If  $P_1$  and  $P_2$  are types, then  $P_1 \rightarrow P_2$  is a type, and  $\lambda x.f_2$  is a canonical inhabitant of it if  $x$  is a variable of type  $P_1$ .
- $\perp$  is a type, and it has no canonical inhabitants.

The type  $P \rightarrow \perp$  will be abbreviated  $\neg P$ .

Also in the language are the following term constructors:

- If  $t$  is a term of type  $(\neg P) \rightarrow P$ , then  $\text{call/cc}(t)$  is a term of type  $P$ .
- If  $t$  is a term of type  $\perp$ , then  $\text{any}^P(t)$  is a term of type  $P$ . The superscript will be omitted when it is deducible from context.
- If  $t$  is a term of type  $P_1 \vee P_2$ , and  $t_1$  is a term of type  $P$  with free variable  $u$  of type  $P_1$ , and  $t_2$  is a term of type  $P$  with free variable  $v$  of type  $P_2$ , then  $\text{decide}(t; u.t_1; v.t_2)$  is a term of type  $P$ .
- If  $t$  is a term of type  $P_1 \wedge P_2$ , and  $t_1$  is a term of type  $P$  with free variables  $u$  of type  $P_1$  and  $v$  of type  $P_2$ , then  $\text{spread}(t; u, v.t_1)$  is a term of type  $P$ .

- If  $f$  is a term of type  $P_1 \rightarrow P_2$ , and  $t$  is a term of type  $P_1$ , then  $\text{apply}(f; t)$  is a term of type  $P_2$ . This term will often be abbreviated  $f(t)$ .

Some abbreviations will be helpful in discussing multiple disjunctions, as  $A_1 \vee \dots \vee A_n$ . We shall parse this as  $A_1 \vee (A_2 \vee (\dots (A_{n-1} \vee A_n)))$ , and we shall use  $\text{in}_i(x)$  as an abbreviation for the appropriate sequence of  $\text{inl}$  and  $\text{inr}$  applications such that if  $x : A_i$ , then  $\text{in}_i(x) : A_1 \vee \dots \vee A_n$ . To generalize the  $\text{decide}$  term constructor to handle multiple disjunctions, assume we have terms  $t_1 \dots t_n$  of type  $P$  such that each  $t_i$  has free variable  $u_i$  of type  $A_i$ , and we have  $t$  of type  $A_1 \vee \dots \vee A_n$ . Then the term  $\text{decide}_n(t; u_1.t_1; \dots; u_n.t_n)$  has type  $P$ , and is considered to be an abbreviation of the appropriate sequence of ordinary  $\text{decide}$  constructions.

### 3 Classical Sequent Proof Terms

For each rule of the classical sequent calculus, we describe how the computational content of the hypothesis of the sequent is transformed to computational content of the conclusion. In general, a sequent  $A_1, \dots, A_m \vdash B_1, \dots, B_n$  will represent the type  $A_1 \rightarrow (A_2 \rightarrow \dots (A_m \rightarrow B_1 \vee \dots \vee B_n))$ , and so its computational content will be a term of that type. Thus, we will assign a variable to each formula on the left of the turnstile, and the inhabitant of the (implicit) disjunction on the right will be constructed from these variables and the term constructors. The computational content of the sequent, then, is the closed  $\lambda$ -term for the function involved.

For all sequent rules but one, we shall see that if we have a constructive proof of each hypothesis sequent, we will also have a constructive proof of the conclusion of the rule. This is true despite the fact that we may have more than one formula on the right side of the turnstile. For, if a sequent  $\Gamma \vdash \Delta$  is constructively provable, given proofs of the formulas in  $\Gamma$  we will be able to determine which formula in  $\Delta$  is proved. This is usually enough to allow us to determine which formula on the left of the conclusion sequent is proved. The exception is the rule of arrow introduction on the left:

$$\frac{\Gamma, A \vdash B, \Delta}{\Gamma \vdash A \rightarrow B, \Delta}$$

Since we may have satisfied the hypothesis sequent by proving something in  $\Delta$ , and we may have used the assumption  $A$ , we cannot be certain that we can still prove anything in  $\Delta$ . Nor can we prove  $A \rightarrow B$ , since we may not have proved  $B$  in the hypothesis sequent. It is only in the computational content of this sequent that the  $\text{call/cc}$  operator appears.

#### Extracting terms from proofs

The rest of this section describes how to construct a term representing the computational content of a sequent proof.

The computational content of an axiom

$$x : A \vdash x : A$$

is the term  $\lambda x.x$ , the identity function. A more general form of this axiom is

$$t_1 : A_1, \dots, t_n : A_n \vdash t_i : A_i$$

and its computational content is the term  $\lambda x_1 \dots \lambda x_n.x_i$ .

Now assume that we are given functions representing the computational content of the sequents which are the premises of each sequent rule. We use these functions to construct the computational content of the conclusion. The notation is extended so that  $f : \Gamma \vdash \Delta$  means that if  $\Gamma = A_1, \dots, A_n$  and  $\Delta = B_1, \dots, B_n$ ,  $f$  is a function with type  $(A_1 \rightarrow (A_2 \rightarrow \dots (A_n \rightarrow (B_1 \vee \dots \vee B_n))))$ . No ambiguity will arise, since the term is read this way only if there is no term explicitly shown for the formula on the right side of the turnstile.

Since sequent rules may involve lists of formulas  $\Gamma = A_1, \dots, A_n$  which are carried from hypothesis to conclusion without modification, we will often abbreviate the list  $g_1 : A_1, \dots, g_n : A_n$  as simply  $g : \Gamma$ . Thus, in the terms below,  $\lambda g.t$  is to be considered an abbreviation for  $\lambda g_1 \dots \lambda g_n.t$ . We also abbreviate

$((f(g_1))g_2)\dots g_n$  as  $f(g)$ , and we may write repeated applications  $(f(x))y$  as  $f(x,y)$  to avoid proliferation of parentheses.

The simplest cases are when there really is only one formula on the right of the turnstile. The computational content corresponds exactly with the standard interpretation. The rules in this case are the following:

$$\frac{f_1 : \Gamma \vdash A \quad f_2 : \Gamma \vdash B}{\lambda g. \text{pair}(f_1(g), f_2(g)) : \Gamma \vdash A \wedge B}$$

$$\frac{f : \Gamma \vdash A}{\lambda g. \text{inl}(f(g)) : \Gamma \vdash A \vee B} \quad \frac{f : \Gamma \vdash B}{\lambda g. \text{inr}(f(g)) : \Gamma \vdash A \vee B}$$

$$\frac{f_1 : \Gamma \vdash A \quad f_2 : \Gamma, B \vdash C}{\lambda g. \lambda x. f_2(g, x(f_1(g))) : \Gamma, A \rightarrow B \vdash C} \quad \frac{f : \Gamma, A \vdash B}{\lambda g. (\lambda x. f(g, x)) : \Gamma \vdash A \rightarrow B}$$

There are also cases in which the rule is not dependent on the number of formulas in  $\Delta$ .

$$\frac{f : \Gamma, A \vdash \Delta}{\lambda g. \lambda x. \text{spread}(x; u, v. f(g, u)) : \Gamma, A \wedge B \vdash \Delta}$$

$$\frac{f : \Gamma, B \vdash \Delta}{\lambda g. \lambda x. \text{spread}(x; u, v. f(g, v)) : \Gamma, A \wedge B \vdash \Delta}$$

$$\frac{f_1 : \Gamma, A \vdash \Delta \quad f_2 : \Gamma, B \vdash \Delta}{\lambda g. \lambda x. \text{decide}(x; u. f_1(g, u); v. f_2(g, v)) : \Gamma, A \vee B \vdash \Delta}$$

Finally, there are the cases in which the fact that there are multiple formulas on the right of some sequent is important. The terms extracted from these cases are more complicated because the conclusion has a disjunction type.

$$\frac{f_1 : \Gamma \vdash A, \Delta \quad f_2 : \Gamma \vdash B, \Delta}{\lambda g. \text{decide}(f_1(g); u. \text{decide}(f_2(g); w. \text{inl}(\text{pair}(u, w)); x. \text{inr}(x)); v. \text{inr}(v)) : \Gamma \vdash A \wedge B, \Delta}$$

$$\frac{f : \Gamma \vdash A, \Delta}{\lambda g. \text{decide}(f(g); u. \text{inl}(\text{inl}(u)); v. \text{inr}(v)) : \Gamma \vdash A \vee B, \Delta}$$

$$\frac{f : \Gamma \vdash B, \Delta}{\lambda g. \text{decide}(f(g); u. \text{inl}(\text{inr}(u)); v. \text{inr}(v)) : \Gamma \vdash A \vee B, \Delta}$$

$$\frac{f_1 : \Gamma \vdash A, \Delta \quad f_2 : \Gamma, B \vdash \Delta}{\lambda g. \lambda x. \text{decide}(f_1(g); u. f_2(g, x(u)); v. v) : \Gamma, A \rightarrow B \vdash \Delta}$$

The case of arrow introduction on the right where there is more than one formula on the right is the only one where assuming the hypothesis is constructive does not imply the conclusion is constructive. It is here we must introduce call/cc.

$$\frac{f : \Gamma, A \vdash B, \Delta}{\lambda g. \text{call/cc}(\lambda h. \text{inl}(\lambda x. \text{decide}((f(g))x; u. u; v. \text{any}(h(\text{inr}(v)))))) : \Gamma \vdash A \rightarrow B, \Delta}$$

To see the intuition behind this term, we consider how it will be used in computation. Suppose  $\Gamma$  is empty, so that we have  $\text{call/cc}(\lambda h. \text{inl}(\lambda x. \text{decide}(f(x); u. u; v. \text{any}(h(\text{inr}(v))))))$  as a closed term of type  $(A \rightarrow B) \vee \Gamma$ . If we use this term in a computation of a value, we must perform a case split, since it is in a disjunction type. When the term is evaluated,  $h$  is bound to the current continuation and we have  $\text{inl}(\lambda x. \text{decide}(f(x); u. u; v. \text{any}(h(\text{inr}(v)))))$  as the term of type  $(A \rightarrow B) \vee \Gamma$ . So, we take the branch

for  $A \rightarrow B$  with the function  $\lambda x. \text{decide}(f(x); u.u; v.\text{any}(h(\text{inr}(v))))$  as our purported inhabitant of the type  $A \rightarrow B$ . Now suppose this function is applied to a term  $a$  of type  $A$ . Applying the function  $f$  from the hypothesis of the sequent to  $a$ , we have an element of  $B \vee \Delta$ . Should this actually be of type  $B$ , then we return that value. If not, we have an inhabitant of  $\Delta$ . We then return the computation to the stage at which we took the case split for  $(A \rightarrow B) \vee \Delta$  by applying the continuation  $h$  to our new evidence for  $(A \rightarrow B) \vee \Delta$ , namely the inhabitant of  $\Delta$  produced by  $f(a)$ . Thus, the continuation operator allows the computation to switch paths when it comes upon further evidence for which of the disjuncts is proved. This description will be made more concrete in the next section when the computation rules are defined.

The structural rules also have term extractions, though their computational content is uninteresting.

$$\frac{f : \Gamma \vdash \Delta}{\lambda g. \text{inr}(f(g)) : \Gamma \vdash A, \Delta} \quad \frac{f : \Gamma \vdash \Delta}{\lambda g. \lambda x. f(g) : \Gamma, A \vdash \Delta}$$

$$\frac{f : \Gamma \vdash A, A, \Delta}{\lambda g. \text{decide}(f(g); u.\text{inl}(u); v.v) : \Gamma \vdash A, \Delta} \quad \frac{f : \Gamma, A, A \vdash \Delta}{\lambda g. \lambda x. f(g, x, x) : \Gamma, A \vdash \Delta}$$

The exchange rules also require treatment, though their computational content is trivial.

$$\frac{f : \Gamma \vdash A_1, \dots, A_i, A_{i+1}, \dots, A_n}{\lambda g. \text{decide}_n(f(g); u_1.\text{in}_1(u_1); \dots; u_i.\text{in}_i(u_i); u_{i+1}.\text{in}_i(u_{i+1}); \dots; u_n.\text{in}_n(u_n)) : \Gamma \vdash A_1, \dots, A_{i+1}, A_i, \dots, A_n}$$

$$\frac{f : A_1, \dots, A_i, A_{i+1}, \dots, A_n \vdash \Delta}{\lambda x_1 \dots \lambda x_{i+1} \lambda x_i \dots \lambda x_n. f(x_1, \dots, x_i, x_{i+1}, \dots, x_n) : A_1, \dots, A_{i+1}, A_i, \dots, A_n \vdash \Delta}$$

## 4 Reduction of terms

In this section we describe a more general term language and present machine rewrite rules for computation in this language. These rewrite rules will define the operational semantics of the language, and it is with respect to this semantics that we will prove that reduction preserves type.

We wish to be able to use these classical proof terms in a general context; for example, we may want to use a classical proof term to represent the propositional logic content of a proof about the integers. To do this, we allow typed nonlogical constants in the term language. We cannot give reduction rules for these constants, since they may represent information about some external theory. We instead require that reductions involving these constants are terminating and type correct, so that constants obey the standard constructive mathematical semantics. For example, if we have a function constant  $f$  of type  $A \rightarrow B$ , if we apply  $f$  to  $a$  of type  $A$ , we are guaranteed that the result  $b$  is of type  $B$ . Furthermore we must be able to treat  $b$  as we would any ordinary member of type  $B$ , so that if  $B = B_0 \vee B_1$ , we must be able to reduce the term  $\text{decide}(b; u.t_1; v.t_2)$ . So we actually require that *all* constants which may appear during reduction (not just the constants present at the beginning) have the property that reductions involving them are terminating and type correct.

Given these assumptions on the behavior of the constants under reduction, we define the program term language to be type-correct terms built from constants and proof terms by any of the term constructors described above, except for `call/cc`. Thus, we know that all instances of `call/cc` in the program are from proof terms, and so we know the context in which they occur.

Since we may be using these proof terms to represent the logical content of a proof in a larger theory, we shall call the result an *applied proof term*. Formally, we define an applied proof term as follows:

- Proof terms are applied proof terms. The type of a proof term is the type corresponding to the formula which was proved.
- Typed constants are applied proof terms.
- Typed variables are applied proof terms.



- If  $t_1$  and  $t_2$  are applied proof terms of type  $P_1$  and  $P_2$  respectively, then  $\text{pair}(t_1, t_2)$  is an applied proof term of type  $P_1 \wedge P_2$ , and  $\text{inl}(t_1)$  and  $\text{inr}(t_2)$  are applied proof terms of type  $P_1 \vee P_2$ .
- If  $x$  is a variable of type  $P_1$  and  $t$  is an applied proof term of type  $P_2$ , then  $\lambda x. t$  is an applied proof term of type  $P_1 \rightarrow P_2$ .
- If  $t$  is an applied proof term of type  $\perp$ , then  $\text{any}^P(t)$  is an applied proof term of type  $P$ . (The superscript will be omitted when it is deducible from context.)
- If  $t$  is an applied proof term of type  $P_1 \vee P_2$ , and  $t_1$  is an applied proof term of type  $P$  with free variable  $u$  of type  $P_1$ , and  $t_2$  is an applied proof term of type  $P$  with free variable  $v$  of type  $P_2$ , then  $\text{decide}(t; u.t_1; v.t_2)$  is an applied proof term of type  $P$ .
- If  $t$  is an applied proof term of type  $P_1 \wedge P_2$ , and  $t_1$  is an applied proof term of type  $P$  with free variables  $u$  of type  $P_1$  and  $v$  of type  $P_2$ , then  $\text{spread}(t; u, v. t_1)$  is an applied proof term of type  $P$ .
- If  $f$  is an applied proof term of type  $P_1 \rightarrow P_2$ , and  $t$  is an applied proof term of type  $P_1$ , then  $\text{apply}(f; t)$  is an applied proof term of type  $P_2$ . This term will often be abbreviated  $f(t)$ .

An applied proof term is called a *program term* if it has no free variables.

Following the example of [3], we present the operational semantics as a set of rules representing the transition function of a term rewriting machine. We use their notion of a continuation point to represent the continuation object; however, since we do not fix an evaluation order, we shall use a more general notion of evaluation context (the analogous "applicative contexts" defined in [3] are defined for a particular evaluation order). An evaluation context is a term with a hole in it, such that the hole is not within the scope of any binding operators (including those in the body of spread and decide terms). A program not in normal form can be split into an evaluation context and a redex. We use the notation  $E[R]$  for such a program, where  $E[]$  is the evaluation and  $R$  is the redex. A continuation point is an evaluation context tagged with  $p$ . If  $E[R]$  is a program of type  $\varphi$  and  $R$  is of type  $\alpha$ , then  $\langle p, E[] \rangle$  is a continuation point which may be applied to any term of type  $\alpha$  to produce a program of type  $\varphi$ .

So, the reduction rules for terms are:

$$\begin{aligned}
E[(\lambda x. t) a] &\mapsto E[t[x := a]] \\
E[\text{spread}(\text{pair}(a, b); u, v. t)] &\mapsto E[t[u := a, v := b]] \\
E[\text{decide}(\text{inl}(a); u. t_1; v. t_2)] &\mapsto E[t_1[u := a]] \\
E[\text{decide}(\text{inr}(b); u. t_1; v. t_2)] &\mapsto E[t_2[v := b]] \\
E[\text{call/cc}(\lambda k. t)] &\mapsto E[t[k := \langle p, E[] \rangle]] \\
E[\langle \langle p, E[] \rangle \rangle a] &\mapsto E_0[a]
\end{aligned}$$

(plus the assumptions described above about reductions involving constants.)

The rule for  $\text{call/cc}$  described above, though not fully general, is sufficient since all of the terms extracted from proofs will have the property that the  $\text{call/cc}$  operator will only be applied to terms of the form  $\lambda k. t$ .

Note that from these rules alone we can conclude that reduction of type correct programs preserves the type of the whole program.

**Theorem 1** *If  $t$  is a well-typed program term of type  $\varphi$ , and  $t$  reduces to  $t'$ , then  $t'$  has type  $\varphi$ .*

*Proof:* The only cases to verify are the rules for reductions of  $\text{call/cc}$  terms, and the application of continuation points. For the case of  $\text{call/cc}$ , if we have  $E[\text{call/cc}(\lambda k. t)]$  is of type  $\varphi$  and  $\text{call/cc}(\lambda k. t)$  has type  $\alpha$ , we have that  $\lambda k. t$  has type  $\neg\alpha \rightarrow \alpha$ , since the program is well-typed. Internally, then, the continuation point  $\langle p, E[] \rangle$  which is substituted for  $k$  will have type  $\alpha \rightarrow \perp$ , but in fact  $E[]$  represents a program of type  $\varphi$  with a hole in it for a term of type  $\alpha$ . (If we had  $A$ -translated the whole program, this difference between the apparent type of the continuation within the program and the actual type of the continuation in computation would not appear.) So, the term  $t[k := \langle p, E[] \rangle]$  has type  $\alpha$ , so the reduction of  $\text{call/cc}(\lambda k. t)$  preserves the type of the whole term.

For the case of the application of continuation points, we showed above that if  $\text{call/cc}(\lambda k.t)$  has type  $\alpha$ , then  $E[]$  represents a program of type  $\varphi$  with a hole in it for a term of type  $\alpha$ . Since its internal type is  $\alpha \rightarrow \perp$ , though, we know that if it is applied, it must be applied to a term  $a$  of type  $\alpha$ , since the program is type correct. This reduction results in  $E[a]$ , which is a program of type  $\varphi$  because  $E[\text{call/cc}(t)]$  was of type  $\varphi$ .  $\square$

## 5 Proof-theoretic interpretation of reduction

### 5.1 Expanded sequent calculus

Since terms in the term language correspond more closely to natural deduction proofs than to sequent proofs, in order to reason about properties of terms as properties of proofs we shall extend the notion of proof with additional rules. The result is a hybrid between natural deduction and sequent systems, with the natural deduction steps appearing as intermediate steps in the sequent proof. Reductions will generally correspond to some preliminary rearranging of the proof, followed by a sequence of cut eliminations. The procedure of substitution corresponds to a sequence of cut elimination steps.

We add to the normal collection of sequent rules (including cut) the rules corresponding to elimination of connectives in natural deduction. We shall then have that every redex corresponds to a natural deduction rule or a cut, although not every cut will correspond to a redex.

A decide redex corresponds to a rule of the form:

$$\frac{\vdash c : A \vee B \quad u : A \vdash f_1 : P \quad v : B \vdash f_2 : P}{\vdash \text{decide}(c; u.f_1; v.f_2) : P}$$

This may also appear as

$$\frac{\vdash c : A, B \quad u : A \vdash f_1 : P \quad v : B \vdash f_2 : P}{\vdash \text{decide}(c; u.f_1; v.f_2) : P}$$

because of the implicit disjunction on the right of the turnstile.<sup>1</sup>

Similarly, a spread which may be reduced corresponds to a rule of the form:

$$\frac{\vdash c : A \wedge B \quad u : A, v : B \vdash f : P}{\vdash \text{spread}(c; u, v.f) : P}$$

Finally, the  $\rightarrow$ -elimination rule corresponds to application:

$$\frac{\vdash a : A \quad \vdash f : A \rightarrow B}{\vdash f(a) : B}$$

If  $f(a)$  is actually a redex, this will reduce to an ordinary cut rule.

### 5.2 The map from terms to proofs

Since we wish to reason about properties of terms by reasoning about the proofs from which they are derived, we need to associate a proof tree with each program term. Furthermore, it is helpful if the proof tree structure reflects the term structure, so we need to augment the sequent proof with some redundant information corresponding to the syntax of the term associated with each sequent rule.

First, we describe the modifications needed to make the original sequent proof correspond to the term structure. The abbreviation  $f(g)$ , used above for repeated applications, requires additional explanation here if each application redex is to correspond to a single cut. In what follows we shall use a notation

$$\frac{g : \Gamma \vdash g : \Gamma \quad f : \Gamma \vdash \Delta}{g : \Gamma \vdash f(g) : \Delta}$$

<sup>1</sup>In the rest of this paper, all of the implicit disjunctions occurring on the right side of the turnstile are now considered to be explicit. This allows us to treat the proof as a pseudo-constructive proof with only one conclusion of each sequent.

for repeated application, where  $g : \Gamma$  is a list  $x_1 : A_1, \dots, x_n : A_n$ , and the above cut is an abbreviation for the sequence of cuts

$$\frac{x_1 : A_1, \dots, x_n : A_n \vdash x_1 : A_1 \quad f : A_1, \dots, A_n \vdash \Delta}{x_1 : A_1, \dots, x_n : A_n, f(x_1) : A_2, \dots, A_n \vdash \Delta} \\ \vdots \\ \frac{x_1 : A_1, \dots, x_n : A_n \vdash x_n : A_n \quad x_1 : A_1, \dots, x_n : A_n, f(x_1, \dots, x_{n-1}) : A_n \vdash \Delta}{x_1 : A_1, \dots, x_n : A_n \vdash f(x_1, \dots, x_n) : \Delta}$$

where after the first rule, the term  $f(x_1)$  has type  $A_2 \rightarrow (A_3 \rightarrow (\dots \rightarrow (A_n \rightarrow \Delta)))$ , etc. This simply describes the repeated application of  $f$  to each of the new variables  $x_1 \dots x_n$ , where each cut corresponds to one application.

We give a few examples of the remaining cases. In general, the structure of the term associated with the conclusion of the rule governs the structure of the proof into which the rule expands. The effect of these transformations is mainly to make function applications and other operations which are implicit in the term explicit in the proof.

For example, the sequent

$$\frac{f : \Gamma, A \vdash \Delta}{\lambda g. \lambda x. \text{spread}(x; u, v. f(g, u)) : \Gamma, A \wedge B \vdash \Delta}$$

becomes

$$\frac{x : A \wedge B \vdash x : A \wedge B \quad \frac{g : \Gamma, u : A, v : B \vdash g : \Gamma, u : A \quad f : \Gamma, A \vdash \Delta}{g : \Gamma, u : A, v : B \vdash (f(g))u : \Delta}}{\lambda g. \lambda x. \text{spread}(x; u, v. (f(g))u) : \Gamma, A \wedge B \vdash \Delta}$$

As another example, the sequent

$$\frac{f : \Gamma \vdash A, \Delta}{\lambda g. \text{decide}(f(g); u. \text{inl}(\text{inl}(u)); v. \text{inr}(v)) : \Gamma \vdash A \vee B, \Delta}$$

becomes

$$\frac{\frac{g : \Gamma \vdash g : \Gamma \quad f : \Gamma \vdash A, \Delta}{g : \Gamma \vdash f(g) : A, \Delta} \quad \frac{u : A \vdash u : A \quad u : A \vdash \text{inl}(u) : A \vee B}{u : A \vdash \text{inl}(\text{inl}(u)) : A \vee B, \Delta} \quad \frac{v : \Delta \vdash v : \Delta}{v : \Delta \vdash \text{inr}(v) : A \vee B, \Delta}}{\lambda g. \text{decide}(f(g); u. \text{inl}(\text{inl}(u)); v. \text{inr}(v)) : \Gamma \vdash A \vee B, \Delta}$$

There is one case where the expansion is not obvious.

$$\frac{f : \Gamma, A \vdash B, \Delta}{\lambda g. \text{call/cc}(\lambda h. \text{inl}(\lambda x. \text{decide}((f(g))x; u. u; v. \text{any}(h(\text{inr}(v))))) : \Gamma \vdash A \rightarrow B, \Delta}$$

The expansion must mirror the structure of the term, but it is not clear how the call/cc and the associated variable  $h$  should be treated. Since the evaluation of a call/cc binds the variable to the current continuation, we shall make this binding explicit by adding an extra cut with a placeholder representing the eventual continuation. We shall call this placeholder  $h$  as well, and represent it as a constant. The last step in the derivation is then:

$$\frac{\vdash h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \quad g : \Gamma, h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \quad \text{inl}(\lambda x. \text{decide}((f(g))x; u. u; v. \text{any}(h(\text{inr}(v))))) : A \rightarrow B, \Delta}{\lambda g. \text{call/cc}(\lambda h. \text{inl}(\lambda x. \text{decide}((f(g))x; u. u; v. \text{any}(h(\text{inr}(v))))) : \Gamma \vdash A \rightarrow B, \Delta}$$

In the derivation of the body of the call/cc, the variable  $h$  remains an assumption along the branch of the proof corresponding to the second branch in the decide:

$$\frac{\frac{h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \vdash h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \quad \frac{v : \Delta \vdash v : \Delta}{v : \Delta \vdash \text{inr}(v) : (A \rightarrow B) \vee \Delta}}{h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp, v : \Delta \vdash h(\text{inr}(v)) : \perp}}{h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp, v : \Delta \vdash \text{any}(h(\text{inr}(v))) : B}$$

The remainder of the derivation of the body of the call/cc is:

$$\frac{\frac{g : \Gamma, x : A \vdash g : \Gamma, x : A \quad f : \Gamma, A \vdash B, \Delta}{g : \Gamma, x : A \vdash (f(g))x : B, \Delta} \quad u : B \vdash u : B \quad (\text{see above})}{\frac{g : \Gamma, x : A, h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \vdash \text{decide}((f(g))x; u; v.\text{any}(h(\text{inr}(v)))) : B}{g : \Gamma, h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \vdash \lambda x. \text{decide}((f(g))x; u; v.\text{any}(h(\text{inr}(v)))) : A \rightarrow B}}{g : \Gamma, h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \vdash \text{inl}(\lambda x. \text{decide}((f(g))x; u; v.\text{any}(h(\text{inr}(v)))) : A \rightarrow B, \Delta}$$

We have described how to augment the sequent proof with the information needed to model computation with it. We must now associate sequent trees with the rest of the term in a similar manner. As previously described, the program term is constructed from terms extracted from sequent proofs, variables, and typed constants, using the usual term constructors. The sequent tree corresponding to the program term is constructed using the type information and the structure of the part of the program term not described by the sequent proof. In essence, we construct a proof of the formula corresponding to the type of the whole program, from assumptions corresponding to the types of the constants.

To describe computation with constants, we associate with each constant a sequent tree of the following form, according to its type. For a constant  $c$  of type  $P$ , the sequent tree is defined inductively as follows. We shall define a map  $[\bullet]$  from sequents to sequent trees, and the tree associated with  $c : P$  will be the tree  $[\vdash P]$ .

$$\begin{aligned} [\Gamma \vdash P] &= \Gamma \vdash P && \text{if } P \text{ is atomic} \\ [\Gamma \vdash A \wedge B] &= \frac{[\Gamma \vdash A] \quad [\Gamma \vdash B]}{\Gamma \vdash A \wedge B} \\ [\Gamma \vdash A \rightarrow B] &= \frac{[\Gamma, A \vdash B]}{\Gamma \vdash A \rightarrow B} \\ [\Gamma \vdash \neg A] &= \frac{\Gamma, A \vdash \perp}{\Gamma \vdash \neg A} \end{aligned}$$

Finally, we define

$$[\Gamma \vdash A \vee B] = \frac{[\Gamma \vdash A]}{\Gamma \vdash A \vee B} \quad \text{or} \quad \frac{[\Gamma \vdash B]}{\Gamma \vdash A \vee B}$$

though we cannot know which until the a cut with  $\Gamma \vdash A \vee B$  is actually eliminated. For example, this case arises when we have a constant  $c$  of type  $A \rightarrow (B \vee C)$ . We cannot know the type of the result until  $c$  has been applied; since we assume that the evaluation of constant functions terminates and is constructively type correct, we will be able to decide what type the result of a particular application of  $c$  actually has.

The above definitions correspond to a tableau proof development except that formulas on the left of the turnstile are not broken down. This is because formulas on the left are arguments to the constant function, and the function must be applied to arguments of the appropriate type.

We now have sequent trees associated with constant terms and with the term extracted from the original sequent proof. We now define a map from terms built from these terms to sequent trees. The map will be denoted  $\{\bullet\}_D$ , where  $D$  is a list of the typed bound variables at any point in the term.

$$\begin{aligned}
\{t_0\}_D &= \text{the augmented sequent proof described above} \\
\{c : P\}_D &= [\vdash P] \\
\{\text{pair}(x, y) : A \wedge B\}_D &= \frac{\{x : A\}_D \quad \{y : B\}_D}{D \vdash \text{pair}(x, y) : A \wedge B} \\
\{\text{inl}(x) : A \vee B\}_D &= \frac{\{x : A\}_D}{D \vdash \text{inl}(x) : A \vee B} \\
\{\text{inr}(x) : A \vee B\}_D &= \frac{\{x : B\}_D}{D \vdash \text{inr}(x) : A \vee B} \\
\{\lambda x. t : A \rightarrow B\}_D &= \frac{\{t : B\}_{x:A, D}}{D \vdash \lambda x. t : A \rightarrow B} \\
\{\text{decide}(d; u.t_1; v.t_2) : P\}_D &= \frac{\{d : A \vee B\}_D \quad \{t_1 : P\}_{u:A, D} \quad \{t_2 : P\}_{v:B, D}}{D \vdash \text{decide}(d; u.t_1; v.t_2) : P} \\
\{\text{spread}(p; u, v.t) : P\}_D &= \frac{\{p : A \wedge B\}_D \quad \{t : P\}_{u:A, v:B, D}}{D \vdash \text{spread}(p; u, v.t) : P} \\
\{\text{apply}(f; x) : B\}_D &= \frac{\{x : A\}_D \quad \{f : A \rightarrow B\}_D}{D \vdash \text{apply}(f; x) : B} \\
\{\text{any}(x) : A\}_D &= \frac{\{x : \perp\}_D}{D \vdash \text{any}(x) : A}
\end{aligned}$$

### 5.3 Reduction as proof transformation

In this section, we describe the correspondence between reduction steps in the program term and cut elimination steps in the proof.

Most reduction steps in the program will correspond to an initial step which creates a cut, followed by a sequence of cut elimination steps. This is because substitution of a term for a variable involves passing through the proof tree until the place where the variable is introduced (as part of an axiom sequent) is found, and substituting a proof for that axiom. Note that this corresponds more closely with the actual complexity of substitution, since the proof tree corresponds closely to the syntactical structure of the term. Most reduction steps follow a pattern of setting up a simple cut and eliminating it, in a sequence of cut elimination steps.

In the remainder of this section, I will describe, for each kind of reduction, how reduction transforms the proof.

Note that the first premise sequent of a redex must be a sequent with no hypotheses, since we do not allow reductions within the scope of a binding operator.

#### 5.3.1 Simple cuts.

Simple cuts arise from the reduction of any of the more complex rules. The elimination of a cut

$$\frac{\vdash a : A \quad x : A \vdash f : B}{\vdash (\lambda x. f) a : B}$$

corresponds to  $\beta$ -reduction, so we must describe how the argument  $a:A$  comes to be substituted for  $x$  in  $f$ . Since the structure of the term corresponds so closely with the structure of the proof tree, in order for the substitution of the argument for the bound variable to take place, the argument must be propagated towards the leaves of the tree to the place where the variable is introduced in an axiom sequent. Observe that this is almost immediate in the terms which come directly from the sequent proof, because these terms are constructed via application from the term(s) representing the hypothesis. For general terms, however, this is not true, and it may take several cut elimination steps for the argument to reach the places where the variable was introduced. So we must pass the cut with  $a:A$  towards the leaves of the sequent tree until it reaches the point where the variable  $x$  was introduced as an axiom.

The procedure is described by induction on the structure of the proof of  $x : A \vdash f : B$ . To eliminate a cut with an axiom sequent, as

$$\frac{\vdash a : A \quad x : A, g : \Gamma \vdash x : A}{\Gamma \vdash A}$$

( $\Gamma$  may of course be empty), replace this cut with

$$\Gamma \vdash a : A.$$

To eliminate a cut with a sequent not an axiom, suppose that  $\Gamma_0 \vdash \Delta_0$  and  $\Gamma_1 \vdash \Delta_1$  are the hypotheses of the sequent rule resulting in  $x : A, g : \Gamma \vdash d : \Delta$ . First, replace the conclusion of the cut rule with  $g : \Gamma \vdash d : \Delta$ . If  $x : A$  is not in either list  $\Gamma_0$  or  $\Gamma_1$ , then we are done. Otherwise, continue the process recursively with the cuts

$$\frac{\vdash a : A \quad \Gamma_0 \vdash \Delta_0}{\Gamma_0 - x : A \vdash \Delta_0} \quad \text{and} \quad \frac{\vdash a : A \quad \Gamma_1 \vdash \Delta_1}{\Gamma_1 - x : A \vdash \Delta_1}$$

(assuming both are applicable, i.e.  $x : A$  is in both  $\Gamma_0$  and  $\Gamma_1$ ).

In other words, we replace the cut

$$\frac{\vdash a : A \quad \frac{\Gamma_0 \vdash \Delta_0 \quad \Gamma_1 \vdash \Delta_1}{x : A, g : \Gamma \vdash d : \Delta}}{\Gamma \vdash \Delta}$$

with

$$\frac{\frac{\vdash a : A \quad \Gamma_0 \vdash \Delta_0}{\Gamma_0 - x : A \vdash \Delta_0} \quad \frac{\vdash a : A \quad \Gamma_1 \vdash \Delta_1}{\Gamma_1 - x : A \vdash \Delta_1}}{g : \Gamma \vdash d : \Delta}$$

and continue the cut elimination process with these new cuts until an axiom is reached.

Note that although we seem to be introducing new sequents  $\Gamma_0 - x : A \vdash \Delta_0$  and  $\Gamma_1 - x : A \vdash \Delta_1$ , in fact these are eliminated at the next stage in the process. In general, the elimination of a cut removes its conclusion sequent and replaces it with another sequent proving the same formula.

### 5.3.2 Reduction of decide

There are two cases, which reduce in essentially the same way.

#### 1. `decide(inl(a);u.t1;v.t2)`

The redex `decide(inl(a);u.t1;v.t2)` corresponds to a proof segment of the form

$$\frac{\vdash a : A \quad \vdash \text{inl}(a) : A \vee B \quad u : A \vdash t_1 : P \quad v : B \vdash t_2 : P}{\vdash \text{decide}(\text{inl}(a);u.t_1;v.t_2) : P}$$

This first reduces to

$$\frac{\vdash a : A \quad u : A \vdash t_1 : P}{\vdash P}$$

and then the process described above for simple cuts is performed.

## 2. $\text{decide}(\text{inr}(b); u.t_1; v.t_2)$

The redex  $\text{decide}(\text{inr}(b); u.t_1; v.t_2)$  reduces similarly to the previous case.

$$\frac{\frac{\vdash b : B}{\vdash \text{inr}(b) : A \vee B} \quad u : A \vdash t_1 : P \quad v : B \vdash t_2 : P}{\vdash \text{decide}(\text{inr}(b); u.t_1; v.t_2) : P}$$

This first reduces to

$$\frac{\vdash b : B \quad v : B \vdash t_2 : P}{\vdash P}$$

Then, this cut is eliminated according to the rules for simple cuts.

### 5.3.3 Reduction of spread

The redex  $\text{spread}(\text{pair}(a, b); u, v.t)$  corresponds to the proof fragment

$$\frac{\frac{\vdash a : A \quad \vdash b : B}{\vdash \text{pair}(a, b) : A \wedge B} \quad u : A, v : B \vdash t : P}{\vdash \text{spread}(\text{pair}(a, b); u, v.t) : P}$$

After reduction, the proof becomes

$$\frac{\vdash b : B \quad \frac{\vdash a : A \quad u : A, v : B \vdash t : P}{v : B \vdash P}}{\vdash P}$$

Then, these simple cuts are eliminated. Note that a sequent  $(B \vdash P)$  has been created, but that it disappears after the cut of  $B$  is eliminated, so that after the reduction we have a proof of

$$\vdash t[u := a, v := b] : P.$$

### 5.3.4 Reduction of applications

An application has the following form:

$$\frac{\vdash a : A \quad \vdash \lambda x.f : A \rightarrow B}{\vdash (\lambda x.f)a : B}$$

Since we have  $\lambda x.f$  as the inhabitant of type  $A \rightarrow B$ , the previous rule must have been an arrow introduction on the right. Hence we must have

$$\frac{\vdash a : A \quad \frac{x : A \vdash f : B}{\vdash \lambda x.f : A \rightarrow B}}{\vdash (\lambda x.f)a : B}$$

This reduces to

$$\frac{\vdash a : A \quad x : A \vdash f : B}{\vdash (\lambda x.f)a : B}$$

without any change in the term itself, and then this cut is eliminated according to the rules for simple cuts above.

### 5.3.5 Continuation proofs

To model the capture and application of continuations in the language, we shall define the notion of a continuation proof. Informally, this will be a partial proof which requires another proof to be complete; in essence, it will be a function from proofs to proofs. For any given proof of a formula  $\phi$ , we will have that any continuation proof arising in the reduction of that proof will, when the continuation proof is applied according to the rules described here, result in a different proof of  $\phi$ .

Under the ordinary intuitionistic interpretation of the logical connectives, a term representing a proof of an implication  $A \rightarrow B$  can be considered a function from proofs of  $A$  to proofs of  $B$ . A continuation proof term is also a function from proofs to proofs; however, it behaves differently when applied within the context of another proof. Suppose we have a proof of a formula  $\varphi$ . A continuation (sub)proof within that proof will appear to be a proof of some formula  $A \rightarrow \perp$ , and the corresponding continuation proof term will appear to have the type  $A \rightarrow \perp$ . Should that term actually be applied, however (in order to produce an inhabitant of  $\perp$ ), what results is not a term of type  $\perp$  within a term of type  $\varphi$ , but rather a new term of type  $\varphi$ , corresponding to a new proof of  $\varphi$ . (Under  $A$ -translation, the  $\perp$  in the proof would have been translated to a  $\varphi$  already.)

The notation we shall use for the continuation proof terms is intended to hint at their role as functions from proofs to proofs. However, we do not wish to consider the internal structure of the continuation proof corresponding to the term. Accordingly, we denote a continuation proof term of type  $A \rightarrow \perp$  as

$$\Lambda x.[]$$

This represents a proof of  $\varphi$  with a “hole” of type  $A$  in it, so that if  $\Lambda x.[]$  is applied to an argument of type  $A$ , the result is a new proof of  $\varphi$ , i.e. a new program of type  $\varphi$ . Describing a continuation proof itself is trickier. Associated with every continuation proof term  $\Lambda x.[]$  is a proof which determines the result when the continuation term is applied. To describe this incomplete proof, we use an ordinary sequent proof, except that at one leaf, instead of an identity axiom, we have the sequent

$$\vdash x : A$$

as an axiom. When the continuation proof is applied, this sequent is replaced with a real proof of  $\vdash A$ , and the term associated with this proof is substituted for  $x$ . (In a language using  $C$  and  $\mathcal{A}$  this would be an ordinary application followed by an abort; since we are using  $\text{call/cc}$ , however, the application and the abort are performed in one step.)

Continuation proofs arise from the reduction of a term  $\text{call/cc}(t)$ . In this system, the only way such a term arises is as the result of an  $\rightarrow$  introduction rule on the right, when there was more than one formula on the right. Thus we know that the form of the proof leading to the term  $\text{call/cc}(t)$  is as described in the previous section. The redex itself appears as the sequent

$$\frac{\vdash h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \quad h : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \vdash \text{inl}(\lambda x. \text{decide}(f(x); u.u; v.\text{any}(h(\text{inr}(v))))) : A \rightarrow B, \Delta}{\vdash \text{call/cc}(\lambda h. \text{inl}(\lambda x. \text{decide}(f(x); u.u; v.\text{any}(h(\text{inr}(v))))) : A \rightarrow B, \Delta} \quad \alpha$$

$\beta$

where  $\alpha$  represents the proof of the other hypotheses (if any) of the rule at this point,  $\beta$  represents the rest of the proof below this rule, and the proof above the  $\text{call/cc}$  is as previously described.

Reduction of this term creates the continuation proof

$$\frac{\vdash d : A \rightarrow B \vee \Delta \quad \alpha}{\beta}$$

where  $d$  is a new variable. The continuation proof term associated with this proof is just  $\Lambda d.[]$ , and it has type  $((A \rightarrow B) \vee \Delta) \rightarrow \perp$ . This continuation proof term is then passed as an argument to the term  $(\lambda h. \text{inl}(\lambda x. \text{decide}(f(x); u.u; v.\text{any}(h(\text{inr}(v)))))$ , where it is substituted for  $h$ . When the dust clears



and the reduction of the original call/cc is complete, the proof structure is

$$\begin{array}{c}
\frac{\frac{\frac{x : A \vdash x : A \quad f : A \vdash B, \Delta}{x : A \vdash f(x) : B, \Delta} \quad u : B \vdash u : B \quad \frac{\frac{\vdash (\lambda d. []) : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \quad v : \Delta \vdash \text{inr}(v) : (A \rightarrow B) \vee \Delta}{v : \Delta \vdash (\lambda d. []) (\text{inr}(v)) : \perp}}{v : \Delta \vdash \text{any}((\lambda d. []) (\text{inr}(v))) : B}}{x : A \vdash \text{decide}(f(x); u.u; v.\text{any}((\lambda d. []) (\text{inr}(v)))) : B}} \\
\frac{\vdash \lambda x. \text{decide}(f(x); u.u; v.\text{any}((\lambda d. []) (\text{inr}(v)))) : A \rightarrow B}{\vdash \text{inl}(\lambda x. \text{decide}(f(x); u.u; v.\text{any}((\lambda d. []) (\text{inr}(v)))) : A \rightarrow B, \Delta} \quad \alpha \\
\beta
\end{array}$$

We now describe how reduction behaves with continuation proof terms. Since continuation terms arise only in the context shown above, if the term is in a redex, it must be that we have taken the right branch of the decide with  $f(a)$  of type  $\Delta$  substituted for  $v$ . So, the proof structure must be

$$\frac{\vdash (\lambda d. []) : ((A \rightarrow B) \vee \Delta) \rightarrow \perp \quad \vdash \text{inr}(f(a)) : (A \rightarrow B) \vee \Delta}{\vdash (\lambda d. []) (\text{inr}(f(a))) : \perp}$$

Since this is a continuation term, the reduction of  $(\lambda d. []) (\text{inr}(f(a)))$  transforms the whole proof to

$$\frac{\vdash \text{inr}(f(a)) : A \rightarrow B \vee \Delta}{\beta} \quad \alpha$$

where the context of this sequent is the context described above, so that the proof reverts to the stage at which the call/cc term was reduced, with the additional information from the proof of  $\vdash f(a) : \Delta$ .

## 6 Conclusion

We have presented a method for extracting programs from classical sequent proofs, which uses the control operator call/cc to represent the classical axiom  $(\neg P \rightarrow P) \rightarrow P$ . As a result, we have a simple proof that reduction of such terms is type preserving, and a proof theoretic framework in which to treat problems of reduction. It is hoped that this will allow the use of tools from proof theory to help solve questions about reduction, such as normalization properties.

## References

- [1] Franco Barbanera and Stefano Berardi. Continuations and simple types: a strong normalization result. manuscript, 1992.
- [2] R. Constable et al. *Implementing Mathematics with The Nuprl Development System*. Prentice-Hall, New Jersey, 1986.
- [3] Matthias Felleisen, Daniel P. Friedman, Eugene Kohlbecker, and Bruce Duba. A syntactic theory of sequential control. *Theoretical Computer Science*, 52:205–237, 1987.
- [4] Jean-Yves Girard. A new constructive logic : classical logic. *Mathematical Structures in Computer Science*, 1:255–296, 1991.
- [5] T. Griffin. A formulas-as-types notion of control. In *Conference Record of the Seventeenth Annual ACM Symposium on Principles of Programming Languages*, 1990.
- [6] Chet Murthy. *Extracting Constructive Content from Classical Proofs*. PhD thesis, Cornell University, Department of Computer Science, 1990.
- [7] J. Underwood. A constructive completeness proof for the intuitionistic propositional calculus. Technical Report 90-1179, Cornell University, 1990.

# Computing the Newtonian Graph

(Extended abstract)

Dexter Kozen\*

kozen@cs.cornell.edu

Kjartan Stefansson\*

stefan@cs.cornell.edu

September 13, 1993

## Abstract

Given a polynomial  $f \in \mathbb{C}[z]$ , it defines a vector field  $N_f(z) = -f(z)/f'(z)$  on  $\mathbb{C}$ . Certain degenerate curves of flow in  $N_f$  give the edges of the Newtonian graph, as defined by [6]. These give a relation between the roots of  $f$  and  $f'$ , much similar to the linear order, when  $f$  has real roots only.

We give an algorithm to compute the Newtonian graph and the basins of attraction in the Newtonian field. The resulting structure can be used to query whether two points in  $\mathbb{C}$  are within the same basin of attraction in  $N_f$ . This gives us an interesting approach to use Newton's method to find all roots of  $f$ , guaranteeing that we converge to a root. This method extends to rational functions and more generally to any functions on  $\mathbb{C}$  whose flow satisfies certain algebraic conditions.

## 1 Introduction

We follow the definitions of Smale [6] and define the *Newtonian vector field* of a polynomial  $f \in \mathbb{C}[z]$  by  $N_f(z) = -\frac{f(z)}{f'(z)}$ . The name is derived from the fact that  $x_{k+1} \leftarrow x_k + N_f(x_k)$  is Newton's method.

---

\*Computer Science Department, Cornell University, Ithaca, NY 14853

The vector field  $N_f$  defines a flow on  $\mathbb{C}$  where the flow comes (almost everywhere) from infinity and converges (almost everywhere) to a root of  $f$ . There are degenerate curves of flow connecting roots of  $f$  and  $f'$ , and the basins of flow split  $\mathbb{C}$  into finitely many regions. These connecting curves will be the edges of our Newtonian graph (to be defined more formally later). This graph has been studied and the types of graphs that arise have been classified [7].

We will give a symbolic algorithm to compute the graph, given a polynomial. Furthermore our algorithm will find the basin boundaries. The output of the algorithm is a structure which can act as an oracle to answer simple questions such as

1. Given  $a, b \in \mathbb{C}$ , are  $a$  and  $b$  in the same basin?
2. Given  $a, b \in \mathbb{C}$ , are  $a$  and  $b$  on the same curve of flow?
3. Given  $a \in \mathbb{C}$  is  $a$  on a basin boundary?
4. Given  $a \in \mathbb{C}$  is  $a$  on a graph edge?

So not only do we get the topology of the graph, we get a method for membership testing for the interesting regions in the field. Such a structure can for instance be used in a "guaranteed" Newton's method, modifying the step size at every point to ensure that we stay within a basin.

We then sketch how to extend the definition of a Newtonian graph for rational functions. We also observe that the resulting fields on  $\mathbb{C}$  satisfy certain algebraic conditions. Given such conditions we can define the graph and compute it.

## 2 The Newtonian Graph

We have defined the Newtonian field of a polynomial. A vector field such as  $N_f$  on  $\mathbb{C}$  defines a flow on  $\mathbb{C}$ . Given  $z \in \mathbb{C}$  the flow through  $z$  is a function  $\phi_z : I \rightarrow \mathbb{C}$ , where  $I \subseteq \mathbb{R}$  is an interval containing zero,  $\phi_z$  differentiable with

$$\begin{aligned} \frac{d\phi_z(t)}{dt} &= N_f(\phi_z(t)) \\ \phi_z(0) &= z. \end{aligned}$$

That is,  $\phi$  parameterizes the flow starting at  $z$  and at every point the speed and direction agrees with the field. An example of flow of  $f$  of degree 4 is given in figure 1.

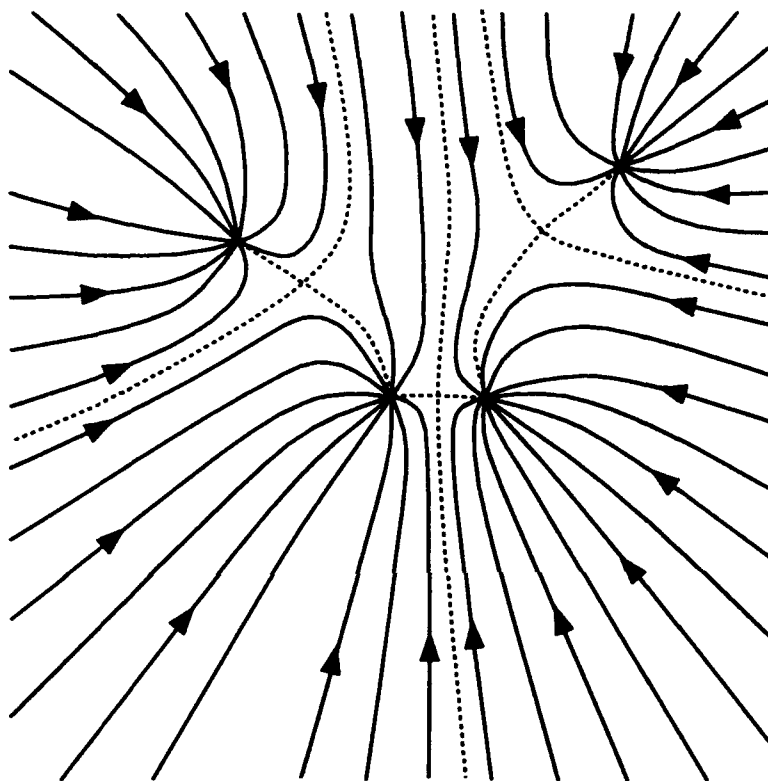


Figure 1: Flow in the Newtonian field of a degree 4 polynomial. Every curve of flow is directed to a root, except the basin boundaries (dotted lines). There is a root of  $f'$  on every basin boundary, and a curve of flow from there to "adjacent" roots (also dotted lines).

The flow exists on all of  $\mathbb{C} \setminus V_{f'}$  (where  $V_{f'} = \{z \in \mathbb{C} : f'(z) = 0\}$ ). The existence and uniqueness follows from the theory of differential equations and the fact that  $N_f$  is a  $C^1$  function on  $\mathbb{C} \setminus V_{f'}$  (see e.g. [3], §8.2 and §8.5).

The following lemma [7] gives us a very important property of the flow:

**Lemma 1** *Let  $f \in \mathbb{C}[z]$ , and  $\phi_z$  be flow through  $z$  in the Newtonian field  $N_f$ . Then  $f$  maps the curve  $\{\phi_z(t) : t \in I\}$  to a straight line pointing to*

the origin. More specifically,

$$f(\phi_z(t)) = e^{-t}f(z).$$

*Proof.* Computing  $\frac{d(\phi_z(t))}{dt}$  using the chain rule gives:

$$\begin{aligned}\frac{df(\phi_z(t))}{dt} &= f'(\phi_z(t))\frac{d\phi_z(t)}{dt} \\ &= f'(\phi_z(t))N_f(\phi_z(t)) \\ &= -f(\phi_z(t)),\end{aligned}$$

which is a differential equation in  $t$  for the function  $\rho(t) = f(\phi_z(t))$ . Given the initial condition,  $\rho(0) = f(z)$ , it has the unique solution  $\rho(t) = e^{-t}f(z)$ , i.e.  $f(\phi_z(t)) = e^{-t}f(z)$ .  $\square$

Using the properties of  $\phi$  one can show the following

**Lemma 2** For every  $z \in \mathbb{C} \setminus (V_f \cup V_{f'})$ ,  $\phi_z$  is defined on a maximum interval containing 0, which is of the following type, for some  $a, b \in \mathbb{R}$ :

1.  $(-\infty, +\infty)$ , and the flow comes in from infinity and goes to a root of  $f$ ;
2.  $(-\infty, a)$  and the flow comes in from infinity and goes to a root of  $f'$ ;
3.  $(a, b)$  and the flow comes in from a root of  $f'$  and goes to another root of  $f'$ ;
4.  $(a, +\infty)$  and the flow comes in from a root of  $f'$  and goes to a root of  $f$ .

*Proof.*  $N_f$  is a  $C^1$  function  $W \rightarrow \mathbb{C}$  where  $W = \mathbb{C} \setminus (V_f \cup V_{f'})$ , and by theorem in §8.5 in [3], all flow must leave any compact set of  $W$ . By Lemma 1, the (maximum) interval of  $\phi_z$  is unbounded upwards iff the flow goes to a root of  $f$ . The same argument shows that the interval is not downward bounded iff the flow comes in from  $\infty$ . Since the flow leaves any compact set of  $W$  the only other limit points are in  $V_{f'}$ .  $\square$

**Definition 3** The *Newtonian Graph* of  $f \in \mathbb{C}[z]$  is the plane graph  $G = (V, E)$  with vertices  $V = V_f \cup V_{f'}$  and directed edges being the curves of flow between vertices, where these exist.  $\square$

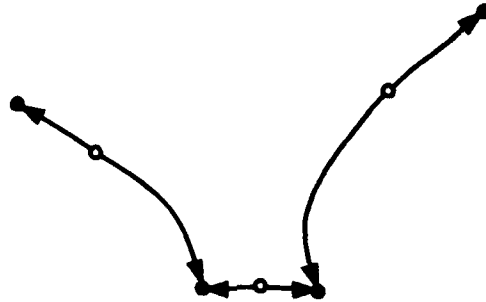


Figure 2: The graph of the field in figure 1. The solid dots are the roots of the polynomial, the hollow ones are the roots of the derivative.

We note that the graph is not just a combinatorial structure, as the edges come with an embedding defined by  $\phi$ .

Figure 2 shows the Newtonian graph of the field shown in figure 1. Another example in figure 3 shows that there can be connections between two roots of  $f'$ .

We observe that under  $f$ , every edge maps onto a straight line segment pointing to 0 in  $\mathbb{C}$ , with at least one endpoint in  $\{f(c) : f'(c) = 0\}$ . This is an immediate consequence of Lemma 1 and the fact that edges are curves of flow. We can conversely look at the pre-images (under  $f$ ) of such line segments, and we get finitely many curves (at most  $n(n-1)$ , since  $f$  is an  $n$  to 1 mapping, and  $f'$  has at most  $n-1$  roots). We will use this observation later, that the graph is contained in the pre-image  $f^{-1}(\{mf(c) : f'(c) = 0, 0 \leq m \leq 1\})$ . Thus the graph has finitely many edges. Furthermore [7] show that the graph is connected and go on to classify the possible types of graphs that can arise.

A basin of attraction is a connected region where the flow comes in from infinity and goes to one particular root of  $f$ . A basin boundary is the boundary of two basins. It is not hard to show that there must be a root of  $f'$  on every basin boundary, because it requires a discontinuity of  $N_f$  for the flow to "split" into two directions, and these are only at the roots of  $f'$ . Also the basin boundaries must be curves of flow themselves, so we conclude that every basin boundary is flow into a root of  $f'$ . In particular this means that basin boundaries are contained in the pre-image  $f^{-1}(\{mf(c) : f'(c) = 0, 1 \leq m\})$ .

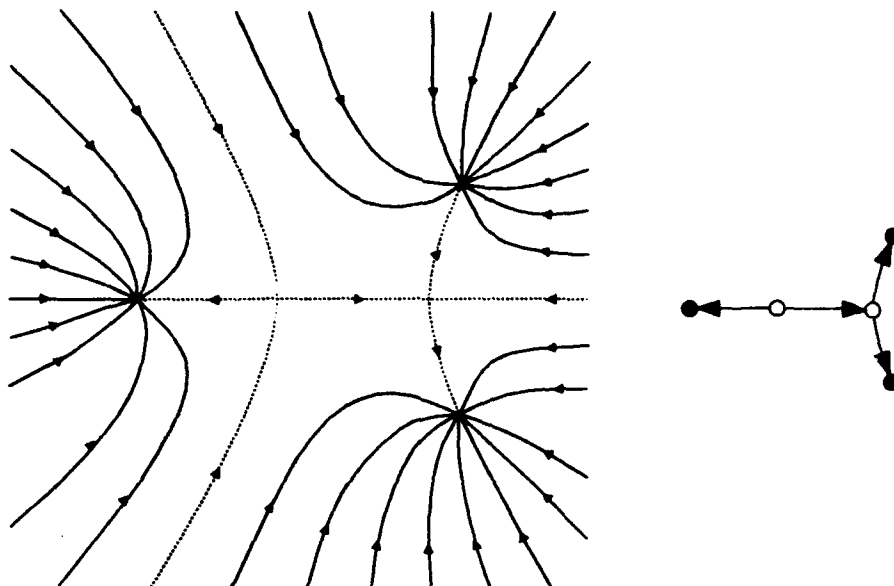


Figure 3: The field and the graph of a degree 3 (real) polynomial where the two derivative roots are linked.

### 3 Computing Basins and Graph Edges

We will give an algorithm to compute the basin boundaries and the edges of the Newtonian graph. But first we need a few preliminaries on cylindric algebraic decomposition.

#### 3.1 Cell Decomposition

We describe cylindric algebraic cell decomposition briefly. For more detailed description, see [2] or [1].

**Definition 4** A *decomposition* of  $\mathbb{R}^k$  is a finite partition  $\{C_i\}_{i \in I}$  such that each  $C_i$  is connected,  $C_i \cap C_j = \emptyset$  if  $i \neq j$  and  $\bigcup_{i \in I} C_i = \mathbb{R}^k$ . For  $k = 1$  such a decomposition is *cylindric* if the each  $C_i$  is either a point or an interval. For  $k > 1$ , the decomposition is cylindric if for all  $r$ ,  $1 \leq r \leq k$ ,  $\{\pi_{1,\dots,r}(C_i) : i \in I\}$  is a cylindric decomposition of  $\mathbb{R}^r$ .  $\square$

**Definition 5** Given polynomial equations,  $f_i(x_1, \dots, x_m) = 0$ ,  $i = 1, \dots, n$ , with  $f_i \in \mathbb{R}[x_1, \dots, x_m]$ , a *Cylindric Algebraic Decomposition* (CAD) of  $\mathbb{R}^m$  is

a data structure  $\mathcal{D}$  with the following properties.

- $\mathcal{D}$  contains a graph, where the nodes correspond to subsets (*cells*) of  $\mathbb{R}^m$ , each cell being homeomorphic to  $\mathbb{R}^d$  for some  $d$ , and the cells are a decomposition of  $\mathbb{R}^m$ .
- For all  $i = 1, \dots, n$ ,  $\text{sign}(f_i)$  is constant on every cell. Each cell is labelled with the signs that the  $f_i$  take on that cell.
- Every node contains an oracle such that given any  $c \in \mathbb{R}^m$ , the oracle can answer if  $c$  is in the subset corresponding to the node.
- Every node contains dimension information, corresponding to the dimension of the cell.
- The edges of the graph correspond to adjacency of the cells in  $\mathbb{R}^m$ , i.e. there is an edge  $(u, v)$  if the subsets that  $u$  and  $v$  represent are adjacent.
- The decomposition is cylindric.

□

Algorithms have been developed to compute (parts of) such a cell decomposition dating back to Tarski in 1948 [8]. Collins [2] has a double exponential algorithm, although it lacks some of the adjacency information. Ben-Or *et al.* [1] developed a parallel algorithm giving the same kind of decomposition (the BKR algorithm), and Kozen and Yap [4] extended that algorithm to obtain the adjacency information as well (here after named the extended BKR algorithm).

We note that due to the cylindric condition and adjacency information, an algorithm computing such a decomposition can be used on a set of polynomials with quantifiers, projecting down the result. If the input is a system of polynomials of the form

$$\begin{aligned} \exists y_1, \dots, y_k : \quad & f_1(x_1, \dots, x_m, y_1, \dots, y_k) = 0 \\ & \vdots \\ & f_n(x_1, \dots, x_m, y_1, \dots, y_k) = 0 \end{aligned}$$

then using CAD on  $\mathbb{R}^{m+k}$  we can project the solution down to  $\mathbb{R}^m$ , by treating the partitions according to  $y_1, \dots, y_k$  as insignificant. The resulting structure



can be used for answering questions of the form: Given  $c \in \mathbb{R}^m$ , does there exist  $y_1, \dots, y_k \in \mathbb{R}^k$  such that  $y_1, \dots, y_k, c$  is a solution to the system?

We note that the order of variables is important with respect to the cylindric condition.

### 3.2 The Algorithm

Recall that every basin boundary and every edge is mapped by  $f$  onto a straight line. Also, the basin boundaries and edges have a root of  $f'$  as a limit. Thus, all these "interesting" curves of flow satisfy, for every  $z$  on the curve,

$$\begin{aligned} \exists c \in \mathbb{C}, m \in \mathbb{R} : f(z) &= mf(c) \\ f'(c) &= 0 \end{aligned} \quad (1)$$

Any point  $z$  on a basin boundary or an edge must satisfy these two conditions. We note that the converse is not true;  $z \in \mathbb{C}$  can be a solution to (1) without being on an edge or a basin boundary.

We proceed in two steps. First we find a decomposition of  $\mathbb{C}$  describing where we have solutions  $z$  to (1). Then we prune that output, because we may get solution curves which do not correspond to basin boundaries or edges.

To find the solutions to (1), we can feed the equations

$$\begin{aligned} f(z) &= mf(c) \\ f'(c) &= 0 \end{aligned} \quad (2)$$

into our favorite cylindric algebraic decomposition algorithm. The resulting structure would be a decomposition of  $\mathbb{R} \times \mathbb{C} \times \mathbb{C}$  describing regions where such  $m, c, z$  exist, along with the dimension of each region and adjacency information. Projecting  $m$  and  $c$ , we get curves in  $\mathbb{C}$  for which there exists a solution to (1).

First let us note that algorithms such as Collins' and the extended BKR do decomposition over the reals. But we can split the equations into a real and imaginary parts, and get a decomposition of  $\mathbb{R}^5 \approx \mathbb{R} \times \mathbb{C} \times \mathbb{C}$ , corresponding to the equations

$$f_R(x, y) = mf_R(c_1, c_2)$$

$$\begin{aligned}
f_I(x, y) &= mf_I(c_1, c_2) \\
f'_R(c_1, c_2) &= 0 \\
f'_I(c_1, c_2) &= 0
\end{aligned}$$

where  $f(x + iy) = f_R(x, y) + if_I(x, y)$  with  $f_R, f_I \in \mathbb{R}[x, y]$ . We get a decomposition on  $\mathbb{R}^5$  which corresponds to a decomposition on  $\mathbb{R} \times \mathbb{C} \times \mathbb{C}$ .

We then project the dimensions corresponding to  $c = c_1 + ic_2$  and again project  $m$ , obtaining a decomposition of  $\mathbb{C}$  corresponding to  $z$  for which there exist  $m$  and  $c$  satisfying equation (2).

This decomposition will contain the basin boundaries and graph edges. These may be partitioned into segments (bounded 1-cells) and 0-cells between such segments. There may be other 1-cells present which are not solutions to the system (introduced by the CAD algorithm to get a finer partition). However, we can always identify the segments which are a solution to this system, because all of them are labelled with the signs of the input polynomials. The curves which are actual solutions to the system (2) will all show  $f(z) = mf(c)$ . Hence a solution curve to the system can be reconstructed by linking such adjacent cells.

But not all solution curves are edges or basin boundaries. The following lemma classifies the types:

**Lemma 6** *The output from the process above contains at most  $O(n^2)$  1-cells which are solutions curves for the input system. They are of the following types:*

1. *Adjacent to two 0-cells, one of which describes a root of  $f'$  and one which describes a root of either  $f$  or  $f'$ ;*
2. *Adjacent only to one 0-cell which describes a root of  $f'$ ;*
3. *Adjacent only to one 0-cell which describes a root of  $f$ .*

*Cells of type 1 are edges of the Newtonian graph, a cell of type 2 is a basin boundary and cells of type 3 are extraneous solutions to the system.*

*Proof.* The only 1-dimensional cells that can be solutions to the system correspond to curves of flow. Then the classification is obvious from the definition of edges and the properties of basin boundaries.

There are at most  $O(n^2)$  solution curves for  $f(z) = mf(c)$  with  $c$  a root of  $f'(c)$  and  $m \in \mathbb{R}$ , because there are at most  $n - 1$  roots of  $f'$  and  $f$  is an  $n$  to 1 mapping.  $\square$

The cells of type 1 and 2 are the ones we are interested in and we must tell these apart from the extraneous cells of type 3.

Since  $f$  is a part of the input, the signs of  $f$  are given on every cell. In particular this allows us to verify if a curve ends at a root of  $f$ .

Depending on which algorithm we use, we may or may not have all the information needed. The extended BKR guarantees that if  $f$  is a part of the input, then the signs of  $f'$  will be provided on each cell. If we don't have this guarantee, we can always add  $f'(z) = 0$  to our input equations and get the same information that way.

At this point we can determine the types of the solution curves. Now it is easy to implement the promised "pruning" step. We simply eliminate all cells of type 3. More precisely, we mark them as parts of the adjacent 2-dimensional cells (which are the basin that this cell lies in).

Now the structure can be used in answering queries. Two points are in the same basin if they are in the same 2-cell or if they are separated only by "fake" 1-cells (of type 3).

## 4 Improvements

Recall we did cylindric decomposition on  $\mathbb{R}^5 \approx \mathbb{R} \times \mathbb{C}^2$  of the equations

$$\begin{aligned} f(z) &= mf(c) \\ f'(c) &= 0 \end{aligned}$$

and projected the solution onto  $\mathbb{C}$ . This can be simplified by defining

$$g(m, z) = \text{Res}_c(f(z) - mf(c), f'(c)),$$

where  $\text{Res}_c$  denotes the univariate resultant of two inputs, considered as polynomials in  $c$ . (Here we view  $f(z) - mf(c)$  and  $f'(c)$  as univariate polynomials in  $\mathbb{C}[z, m][c]$ ).

Then  $g$  has the property that  $g(m, z) = 0$  iff  $\exists c \in \mathbb{C} : f(z) - mf(c) = 0 = f'(c)$ . Hence, a decomposition of  $\mathbb{R} \times \mathbb{C}$  with respect to  $g$  is the same as

the projection of the decomposition of  $\mathbf{R} \times \mathbf{C} \times \mathbf{C}$  with respect to the original two equations.

The only thing we must be aware of is how to obtain the necessary signs of  $f$  and  $f'$  on cells, in order to identify and link up solution curves and prune off the redundant ones. One way would be to add the equation  $f(z) = 0$  (and  $f'(z) = 0$ , if we are not using the extended BKR), and do a decomposition with respect to  $f$  ( $f'$ ) and  $g$ . This is already an improvement in terms of dimensions, since we are only working with 3 real variables ( $x = \text{Re}(z), y = \text{Im}(z)$  and  $m$ ) instead of 5 before.

The asymptotic complexity remains the same, but the constants are clearly much better. The extended BKR gives an NC circuit of depth  $2^{O(d^2)} \log^{O(d)} n$  where  $d$  is the number of variables and  $n$  is the maximum of either the number of polynomials or their degrees. In our case the circuit will be of depth  $O(\log^{O(1)} n)$  where  $n$  is the degree of the input polynomial  $f$ .

## 5 Applications

The Newtonian graph is of its own interest, as it describes the arrangement of the roots of  $f$  and  $f'$ . Our computation gives a complete topological information of both the graph and the basins of the Newtonian field.

The relation to Newton's method gives an interesting method of approximating all the roots of  $f$  simultaneously, guaranteeing convergence. If we start with a point  $z_0$  in a basin, we can apply modified Newton's method, where the iteration  $z_{k+1} \leftarrow z_k + N_f(z_k)$  is replaced by:

```

 $\alpha \leftarrow 1,$ 
repeat
   $z_{k+1} \leftarrow z_k + \alpha N_f(z_k)$ 
   $\alpha \leftarrow \alpha/2$ 
until ( $z_{k+1}$  is in the same basin as  $z_k$ )

```

*I.e.* we scale down the step in order to ensure that we stay within the same basin. Here we use our pre-computed structure of basins to determine if two points are in the same basin. If we furthermore require that a progress is made at each step, (*i.e.* that  $|f(z_{k+1})| < |f(z_k)|$ ), then we are guaranteed that the sequence  $\{z_k\}$  will eventually converge to the root in the basin of  $z_0$ .

We remark that this does not guarantee quadratic convergence everywhere, it only ensures that the method will converge.

## 6 More General Newtonian Graphs

In this extended abstract, we will briefly describe how the Newtonian graph and its computation extend to more general vector fields on  $\mathbb{C}$ .

The definition of a Newtonian graph (Definition 3) only uses the fact the function  $f : W \rightarrow \mathbb{C}$  is  $C^2$  on an open subset  $W \subseteq \mathbb{C}$ , which makes  $N_f(z) = -f(z)/f'(z)$  a  $C^1$  vector field on  $W$ . Lemma 1 still holds, but Lemma 6 now allows curves parameterized  $(-\infty, +\infty)$  coming in from either infinity or a pole of  $f$  and going to either infinity or a root of  $f$ .

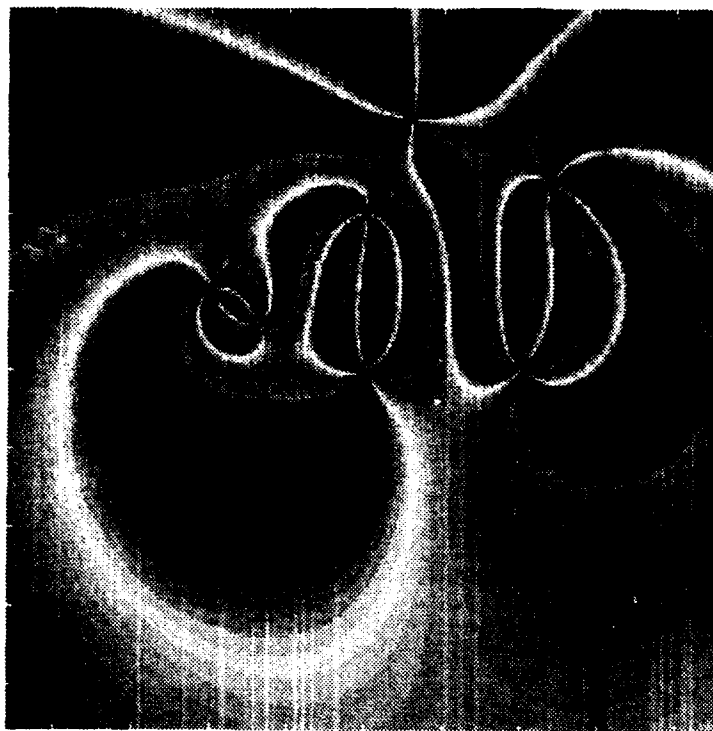


Figure 4: Flow in the Newtonian field of a rational function with three roots and four poles. Curves of fixed color indicate curves of flow.

In particular, most of the same observations hold for rational functions. The function  $f$  maps the curves of flow onto straight line segments pointing to the origin. The natural definition of a Newtonian graph for the rational function is the graph whose edges are the directed curves from poles or infinity to roots of  $f'$ ; between the roots of  $f'$  and from roots of  $f'$  to roots of  $f$  or to infinity. This differs only slightly from Smal's definition [6] in that for polynomials we now count the basin boundaries as a part of the graph. A nice property of the graph is that it is symmetric in poles and roots, i.e. the graphs of  $N_{p/q}$  and  $N_{q/p}$  are identical, except the directions of the edges are reversed.

Write  $f(z) = p(z)/q(z)$  with  $p, q \in \mathbb{C}[z]$  with  $\gcd(p, q) = 1$ . If  $c$  is a root of  $f'(c) = 0$  then  $f$  maps any edge into  $c$  onto a ray  $\{mf(c) : 0 \leq m \leq 1\}$ . Again we can consider a system of equations

$$\begin{aligned} \exists c \in \mathbb{C}, m \in \mathbb{R} : f(z) &= mf(c) \\ f'(c) &= 0 \end{aligned}$$

which now is equivalent to the system of polynomials

$$\begin{aligned} \exists c \in \mathbb{C}, m \in \mathbb{R} : p(z)q(c) &= mq(z)p(c) \\ p'(c)q(c) - p(c)q'(c) &= 0. \end{aligned}$$

This we can solve with algebraic decomposition as before and determine the actual solution curves which are edges. As before we can reduce the number of variables by using resultants. Let

$$g(m, z) = \text{Res}_c(p(z)q(c)mq(z)p(c), p'(c)q(c) - p(c)q'(c)).$$

Then the previous system is equivalent to  $\exists m g(m, z) = 0$ . As before we have devised an  $NC$  algorithm to compute the Newtonian graph.

The key property used in the computation is that the flow satisfies  $f(\phi_z(t)) = e^{-t}f(z)$ . In the case of a rational function  $f(z) = p(z)/q(z)$  this translated into the polynomial equation

$$p(\phi_z(t))q(z) - e^{-t}p(z)q(\phi_z(t)) = 0.$$

This equation also implicitly defines the flow  $\phi_z$  satisfying  $\phi_z(0) = z$ .

In more generality, consider any flow  $\phi_z$  defined on  $W \subseteq \mathbb{C}$ , where  $W$  is almost all of  $\mathbb{C}$ . Assume for some polynomial  $g$  with complex coefficients and for all  $z \in W$ ,  $\phi_z(0) = z$  and  $g(\phi_z(t), e^t, z) = 0$ . Then we have  $Dg(\phi_z(t), e^t, z) = 0$ , i.e.

$$D_1g(\phi_z(t), e^t, z)\phi'_z(t) + D_2g(\phi_z(t), e^t, z) = 0$$

which is equivalent to

$$\phi'_z(t) = -D_2g(\phi_z(t), e^t, z)/D_1g(\phi_z(t), e^t, z). \quad (3)$$

If a basin of attraction is the area where all the flow goes from one pole to one root then as before there is a discontinuity in the field somewhere along every basin boundary. In particular,  $\phi'$  will be undefined at such points. Equation (3) shows that  $\phi'_z(t)$  extends continuously to all of  $\mathbb{C}$  except the points where  $D_1g(\phi_z(t), e^t, z) = 0$  and  $D_2g(\phi_z(t), e^t, z) \neq 0$ . These points can be computed as being the  $x \in \mathbb{C}$  for which there exist  $m, m' \in \mathbb{R}$  and  $z, x' \in \mathbb{C}$  with

$$\begin{aligned} g(x, m, z) &= 0 \\ g(x', m', z) &= 0 \\ D_1g(x', m', z) &= 0. \end{aligned}$$

The first two conditions force  $x$  and  $x'$  to be on the same curve whereas the third condition places  $x'$  at a discontinuity of the field. It is easy to verify that for rational functions we get the same equations as before.

Again we can do CAD on  $\mathbb{R}^5$  and project down for the  $x$  variable to obtain the solution curves, which then are the edges of the Newtonian graph.

## 7 Acknowledgments

This work was supported in part by the National Science Foundation and in part by the United States Army Research Office through the Army Center of Excellence for Symbolic Methods in Algorithmic Mathematics (AC-SyAM), Mathematical Sciences Institute of Cornell University under contract DAAL03-91-C-0027.

## References

- [1] Michael Ben-Or, Dexter Kozen and John Reif *The Complexity of Elementary Algebra and Geometry* Journal of Computer and System Sciences, Vol. 32 No. 2, April 1986
- [2] G. E. Collins, *Quantifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition*, Lecture Notes in Computer Science, vol. 33, pp. 134-183, Springer-Verlag, Berlin Heidelberg New York 1975.
- [3] Morris W. Hirsch and Stephen Smale *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, 1974
- [4] Dexter Kozen, Chee-Keng Yap *Algebraic Cell Decomposition in NC*, Proceedings of the 26th Annual Symposium on Foundations of Computer Science (FOCS) 1985, pp. 515-521
- [5] Morris Marden *The geometry of the Zeroes of a Polynomial in a Complex Variable*, American Mathematical Society, New York, 1966
- [6] Stephen Smale *On the Efficiency of Algorithms of Analysis*, Bulletin of the American Mathematical Society, Vol. 13 No. 2, pp. 87-122, October 1985
- [7] Michael Shub, David Tischler and Robert F. William *The Newtonian Graph of a Complex Polynomial*, SIAM J. Math. Anal., Vol. 19, No. 1, January 1988
- [8] Alfred Tarski, *A Decision Method for Elementary Algebra and Geometry*, Univ. of Calif. Press Berkeley, 1948; 2nd ed. 1951.



# **IMPLEMENTING MIXED CHAINING IN A CLASSIFICATION TYPE EXPERT SYSTEM.**

**Andrew W. Harrell**

**U.S. Army Engineer Waterways Experiment Station,  
Vicksburg, MS 39180**

**ABSTRACT.** Because of the general to specific nature of the backward ordered reasoning (from goals to input data) in some expert systems, it is hard to organize sets of rules that lead to multiple goals. In classification type expert systems, in particular, it is often difficult to organize the rules. Normally it is desired that under all circumstances they ask for all the information required. They should then conclude with a report which contains all the conclusions the system should reach in this situation. In this project, an auxiliary computer program was written to topologically sort the 120 rules in the knowledge base of an expert system. The conclusions of the rules were used as the means by which to define a partial order of the logic flow in the knowledge base.

Key words - Expert System, mixed chaining, knowledge base, topological sort.

**INTRODUCTION.** Generic categories of expert systems applications include decision management, diagnosis /troubleshooting (determining malfunctions from symptoms and other observable facts), classification and interpretation of situations (concluding situation descriptions from the data and facts encountered), planning and scheduling analysis, manufacturing design, configuring objects under constraints, instruction and intelligent documentation, configuration design, and process control (programs to govern the overall behavior of systems).

In 1989 the US Army Engineer Waterways Experiment Station (WES) established a research and development work unit within the Civil Works Research and Development Program's Flood Control Channels Budget Package entitled "Gravel and Boulder Rivers" (#32553). This effort has two major goals: the first being to develop an understanding of the physical sedimentary processes in rivers and streams, the second being to develop a conceptual model of these processes. An initial stream reach inventory form was developed and validated during 1989-1991. Based on the data gained by a nationwide inventory conducted by MCI Consulting Engineers, INC. for WES, a lack of understanding of and data for boulder/gravel systems became apparent. Work was done to:

- a. Establish a systemic procedure for collecting and

analyzing geomorphological, geometric, hydraulic, and sedimentary data using a stream reach inventory process.

b. Identify sediment sources and deposition zones.

c. Identify channel bed and bank forms which are hydraulic influencers.

d. Relate channel processes to channel features and link the sedimentation patterns to river engineering factors.

Efforts were conducted to develop technical guidance documents for use by District personnel. The end product was envisioned to be the basis for uniform data collection methods for boulder/gravel river systems. As a result of work conducted on this project in 1992 an existing set of separate stream bed channel flow rules was organized into a classification type computer expert system by the author using the methodology explained in this report.

This paper describes a knowledge based expert system entitled CHANNEL-FIX. The program is intended to serve the hydraulic engineer as a Boulder/Gravel River sedimentation analysis tool. CHANNEL-FIX provides guidance in the fluvial geomorphic processes occurring in a Boulder/Gravel river reach linked to 5 of the 6 stream channel design variables.

In terms of the description of how the rules are implemented, the scope of this study is limited to a particular version of the expert system software shell used (Level5 ver 1.1 for the MacIntosh). The general procedures to be explained in this report are applicable to this type of software expert system shell in general but the specific syntax and grammar of the rules in the knowledge base and system specific functions will be different for other shells.

Since the stream bed flow expert system program falls under the type of expert system used for classification and interpretation of situations some of the specific characteristics for expert systems in these areas will be briefly described below.

Classification expert systems help the user to choose products, procedures, or processes from a large or complex set of alternative possibilities. These programs identify a hypothesis based on the pattern of data that the user enters in response to a series of questions. Since the questions are asked in response to a set of presupplied hypotheses (that is they are framed and scheduled from the general to the specific) these systems are basically backward-chaining. However, as will be explained below, in some situations the information that accumulates as the data is entered may influence the order in which the questions should be asked. To take account of this the knowledge base and inferencing strategy also need to continue accumulating information even after each partial conclusion is reached. This

may require the expert system to start back through the rules again or iterate repeatedly by forward chaining through the rule sets several times.

The following short glossary defines some of basic terms that will be used in explaining the problems that arose in designing the knowledge base for the expert system shell:

### Terminology

A short list of some basic expert system terminology is listed below <sup>1</sup>:

**Attribute** -- Defines the qualities or values contained in a class and the type of information that make up a class. For example, the class car can have the attributes "type of engine" and "top speed".

**Attribute value** -- An actual number or confidence factor representing the degree of certainty with which a factor is known.

**Backward-Chaining** -- An inferencing strategy that is structured from the general to the specific. That is, it starts with a desired goal or objective and proceeds backwards along a series of deductive reasonings while it attempts to collect the hypotheses required to be able to conclude the goal. This process continues until the goal is reached and it then displays its conclusion. (See following sections for a more complete explanation and an example.

**Class** -- Defines the structure (in terms of its attributes) and behavior (in terms of its associated methods and procedures) of an object. When it becomes an instance, it then holds the actual data values of a particular realization of this type of object in the knowledge base. For example: a class called human beings might have attributes related to the parts that differentiate our physical beings and categories such as those related to its our mental and spiritual capacities. Some of the associated methods and procedures of this class could be thinking, talking, walking. It can be considered as a subclass of another class such as the class of living beings. The author and the reader are both specific instances of a human being object.

**Antecedent** -- The IF part of a conditional statement (synonymous with the term hypothesis in what follows).

**Consequent** -- The THEN part of a conditional statement (synonymous with the term conclusion in what follows).

---

See also, the Level5 object for Windows Users guide, Clips users manual, and a guide to expert systems by Waterman all of which are listed in the bibliography at the end of the report.

**Control Rule --** A rule in the knowledge base that controls the order in which data is assimilated into the knowledge base.

**Goal ---** A top-level consequent of the rules in the knowledge base toward which Backward-Chaining may be directed. (It is a hypothesis that the program will try to determine if some group of rules can be instantiated together to satisfy)

**Inference Mechanism --** The component of the expert system shell responsible for using the rules in the knowledge base to derive new facts from known information.

**Instance or Instantiation --** Specific occurrence of an object. An object consists of its class structure, which defines its attributes and behavior and its instances, which hold the actual values of the object. An instance of the class human beings mentioned above would refer to an individual person, such as the reader of this report.

**Knowledge Tree ---** A graph showing the logic and data flow connections between rules and facts in the knowledge base. A knowledge tree presents a graphical representation of the complete structure of the knowledge base.

**Method ---** A procedure stored in an object's class structure that can determine an attribute's value when it is needed in the program, referenced in its class, or required to execute a series of procedures because another value in the program changes. "When needed methods" are executed during backward chaining to determine an attribute's value. "When changed methods" implement a procedure when a given attribute changes.

**Node ---** A vertex or point in the knowledge tree connecting the antecedents and consequents of rules in the knowledge base. In most conventions the nodes are the rules and the antecedents and consequents are the edges between the nodes or vertices.

**Object --** General term for a programming entity that has a record type data structure along with attribute values and procedures that enable it to represent something concrete or abstract. It can be contrasted with other programming entities such as facts, rules, procedures, or methods. An object's structure is defined by its class and attribute definitions. A class declaration is a data template involved in representing knowledge which defines the structure of an object. For example, in the class "human being" mentioned above some of the attribute slots might be size, weight, hair color, and so forth.

**Expert System --** A computer program that represents and uses expert human knowledge to attain high levels of performance in a problem area. An expert system has two basic components: a knowledge base which contains the information (facts, rules, and methods) found in the subject area of the problem area being represented, and an inference engine or mechanisms that make use

of the knowledge base (by scheduling and interpreting the facts, rules, and methods) to make conclusions and decisions and solve problems that would normally take a human expert more effort.

**Expert System Shell --** The interactive programming environment on the computer into which the user enters information, rules, and goals and which compiles the knowledge base, then runs the resulting expert system program.

**Forward-Chaining --** Forward-chaining reasoning is an inferencing strategy in which the questions are structured from the specific to the general. That is, it starts with user supplied or known facts or data and concludes new facts about the situation based on the information found in the knowledge base. This process will continue until no further conclusions can be reached from the user supplied or initial data (using the rules and methods contained in the knowledge base). (See following sections for a more complete explanation and an example).

**Vertex ---** Same as node. (See above)

## EXAMPLE OF HOW THE RULE-BASED SYSTEM CAN CLASSIFY THE PLANIFORM STABILITY OF A REACH IN AN ACTUAL STREAMBED

The WES CHANNEL-FIX rule based system contains about 80 rules, 101 facts in 950 lines of computer code for a MacIntosh personal computer. Once the program is started on the computer, a screen appears which explains the system which is driven by graphical menus and buttons. The user enters information into the program by either clicking buttons on the screen with a cursor directed by a keyboard mouse or by typing text from the keyboard in order to answer the questions that appear on the screen. Certain menu choices or questions in the program are preceded by explanatory pictures on the screen. These pictures give the user a graphical explanation of some of the menu choices that are displayed. Also, when the explain button appears above the question area on the screen window, the program will display explanatory text when this button is clicked. When the system has asked all the questions that it needs to determine which rules and facts may be applicable to the situation the session will occur a summary of all conclusions and determinations will be printed out on the screen and saved in a file in the program's directory work area. At any time during the series of questions that the program makes a partial conclusion the user may click on the explain button to see displayed which rules and facts were used to make that particular conclusion.

As an example of the steps involved in using the program we will display the questions and determinations for a session in which the user enters the information for a reach in the North Fork Licking river. Normally a reach of the river would be determined from the data from several cross-sections at the site. After the initial screen appears the next step in the program is normally initiated by clicking on the continue button that appears above the question area in the program's screen window.

### a. SAMPLE PROGRAM RUN CROSS SECTION #1 NORTH FORK LICKING RIVER

(1) The program asks for the name of the river which the user enters as **North Fork Licking** in this case.

(2) The program then asks for the type of bar that is present in this reach of the river survey. In this case the user responds: **Point Bar**.

(3) The program then asks for the active channel width<sup>2</sup> in feet. This is entered as **75**.

(4) The program then asks for the slope of the river bed and the water surface slope at this point in the stream. The answers entered in this case are: **.01** and **.02**.

---

<sup>2</sup> See the reference Harrell[1993] for the definition of the hydrologic terms used in this example.

(5) The program then asks the user whether fines are present on the bar surface. The answer is Yes in this example.

(6) The program then asks whether large clasts are in direct contact at this point in the reach. The answer is No in this example.

(7) The program then asks whether imbrication is present. The answer given is Yes.

(8) The program then asks whether you can identify the evidence of fresh scour on the outside bank. The answer given is Yes.

(9) The program then asks whether there are fresh deposits on the bar. The answer given is Yes.

(10) The program then asks what is the average depth of the active channel. The answer given is 2 ft.

(11) The program then asks whether there is fresh scour on the bar. The answer given is No.

(12) The program then asks whether there are diffuse gravel sheets. The answer given is Yes.

(13) The program then asks whether the Main Channel is increasing, stable, or decreasing. The answer given is increasing.

The program then concludes the session and prints out a screen displaying all the conclusions reached. This information is shown below:

Based on your description of this reach of North Fork Licking River, the following conclusions can be drawn:

The sedimentary structure of the bar is Matrix Gravel  
Large clast are not typically in direct contact in Matrix gravels.

The matrix consists of 30% or more sediment finer than fine gravel. Fluvial action will rapidly entrain the matrix sediment reducing the stability of the gravel clast.

This erosional process occurs at mean flow or higher. Field data indicates the even burial of clast to 75% does not increase stability.

Tractive or shear stress produced by mean flow will entrain the matrix finer grain sediment. The lack of clast interlocking that is present in framework gravel reduces the stability although there is a high per cent of fine grained matrix material.

Matrix gravel units appear to be a grouping clusters. This lack of stability and high erodibility factor leads the assignment of a stability rating of 4.

The relative stability of the bar (from 1 to 4) is: 4.00  
The Active Channel Width is probably increasing

The slope is:increasing  
The meander pattern is:decreasing  
The stability of the channel is:decreasing  
Conclusion: The channel is migrating to the outside  
The bank is providing transported sediment.  
The sediment transport is:increasing  
The Main Channel Depth is stable  
The stability of the planiform is: + 2.00  
where +1 = change/increase  
0 = neutral  
-1 = change/decrease

The bank is eroding: True  
The bar is eroding: False  
The bar is migrating: False  
The bar is growing: True

Report of conclusion for North Fork Licking River complete.  
\*\*\*End of session\*\*\*

Note, that in this example the program had to ask the question whether the Main Channel was increasing, stable, or decreasing. For another set of reach information it is possible that the rule-based system would have been able to determine this from information already entered. In general, there are not enough rules to determine all the conclusions that may be required in order to proceed completely with any given set of facts. The program will then request the user to supply the answer to the missing information. The purpose of sorting the rules as explained earlier in the report in terms of the information required in the hypotheses of each rule is that the program will in all cases be able to proceed in a single program run in a manner which extracts all the information required to make all possible determinations that the rule-base will permit.

If we examine the stream bed flow program we see that it is a rule-based system which collects or makes a report of a series of conclusions, not just one. Therefore it does not fall in the area of backward goal searching diagnostic programs in which the questions are structured from the general to the specific. It is a forward chaining rule-based system in which the information is accumulated by asking a series of questions which are structured from the specific to the general.

The order of the goals in the program was restructured in order to make it consider all the rules in a single program run.



An abbreviated forward chaining flow chart for stream bed flow rule system as it now exists is illustrated below:

step 1

Determine the river name  
Determine the type of bar  
Determine the active channel width

step 2

Determine bar composition

step 3

Determine bar stability  
Determine if the bank and/or bar is eroding or not  
Determine channel depth  
Determine channel slope  
Determine if the bar is migrating or not

step 4

Determine if the active channel width is changing  
Determine if the main channel depth is changing  
Determine if the bar and/or the bank is providing  
transported sediment  
Determine if the bar is eroding faster than the bank

step 5

Draw conclusions about the present state of the width, such  
as a point bar is forming  
Draw conclusions about the effect of the bar on  
the active channel width  
Draw conclusion if deposition is occurring on the inside of  
the bend  
Draw conclusions about the affect of increasing channel width on  
slope  
Draw conclusions about the affect of diffuse gravel sheets on  
slope  
Draw conclusions about the affect of slope on channel stability

step 6

file a report of all the information entered  
and conclusions reached

The reason for this organization comes from both the way the software is written and the type of knowledge base that we want to create. The rest of the report will further elaborate on the organization and explain how it was arrived at.

The version of LEVEL5 that was used for this study is a backward chaining (goal driven) PROLOG<sup>3</sup> <sup>4</sup>type expert system shell based on predicate calculus. It provides a good graphical user interface, built-in database search predicates, and some object-oriented features. For classification problems which are data-driven and for which you need to record everything that can be determined about the situation, a forward-chaining LISP<sup>5</sup> <sup>6</sup> or CLIPS<sup>7</sup> type system with more object-oriented features is better.

---

<sup>3</sup> "A logic programming language based on predicate calculus", Barker, 1988.

<sup>4</sup> The book "PROLOG, Programming for Artificial Intelligence", by Ivan Bratko, Addison-Wesley Publishing Co, 1986 contains a well-written and readable guide to understanding how Prolog type expert system programs work. See also, "Logic Programming and Knowledge Engineering" by Tore Amble, Addison Wesley Publishing Co., 1987.

<sup>5</sup> LISP - "A programming language well suited for list processing and symbolic manipulation. It is currently the most popular AI language in the United States", Barker, op.cit.

<sup>6</sup> LISP 3rd ed., by Patrick Henry Winston and Berthold Horn, Addison Wesley Publishing Co, 1989.

<sup>7</sup> CLIPS User's Guide, by Joseph C. Giarratono, NASA Lyndon B. Johnson Space Center, Information Systems Directorate, Software Technology Branch, 1991.

## **A Brief Description of the Way Forward and Backward Chaining Works**

In forward chaining, the inference mechanism starts by evaluating the first rule in the knowledge base. If the antecedent of that first rule is true, then the consequent of the rule is used to search for a conditional with an antecedent identical to the previous consequent. This forward chaining continues until the system is unable to match a consequent with an antecedent. Because the system reasons from the information or data provided, this form of processing is said to be **data driven**. The two rules below will serve as a demonstration of this process:

IF A IS true, THEN B IS true  
IF B IS true, THEN C IS true

a. The following steps define how forward chaining could be applied to the rules above:

(1) If "A is true" is known, the inference mechanism will prove "B is true" by modens ponens<sup>8</sup>.

(2) The system then searches forward for a rule that has an antecedent that matches the consequent "B is true". A match is found in the second rule.

(3) Again the law for modens ponens is used to prove that "C is true". Since no further rules can be found with antecedents that match consequents, the system will offer "C is true" as its conclusion.

In backward chaining, the inference mechanism starts with a goal and seeks to find a rule with that goal as its consequent. It then verifies whether or not that rule can be derived from another rule by finding another rule whose consequent matches its antecedent. This process of backward chaining continues until a rule is found that has an independent antecedent. Thus, backward chaining is actually **goal driven** in its problem solving strategy.

The example below demonstrates the implementation of this concept:

Goal statement

---

<sup>8</sup> "A rule of inference that states: IF A implies B and A is known to be true, then B is true.

D IS true

#### Production rules

RULE 1

IF A IS true  
THEN B IS true

RULE 2

IF B IS true  
THEN C IS true

RULE 3

IF C IS true  
THEN D IS true

b. The first statement in the example above, "D IS true," is the goal for this knowledge base. The following steps explain how LEVEL5's inference mechanism backward chains to prove this goal:

(1) The system begins by searching for a rule with the goal "D IS true" as its consequent. Since Rule 3 satisfies this condition, the program backward chains to check if the antecedent "C IS true" can be derived from another rule.

(2) It is discovered that Rule 2 does, in fact, have a consequent that matches the antecedent of Rule 3. The program will now test to see if the antecedent of Rule 2, "B IS true", can be derived from another conditional.

(3) Rule 1 has a consequent that matches the hypothesis in Rule 2. LEVEL5 searches once more for other supporting rules. Since none can be found, the program asks the user:

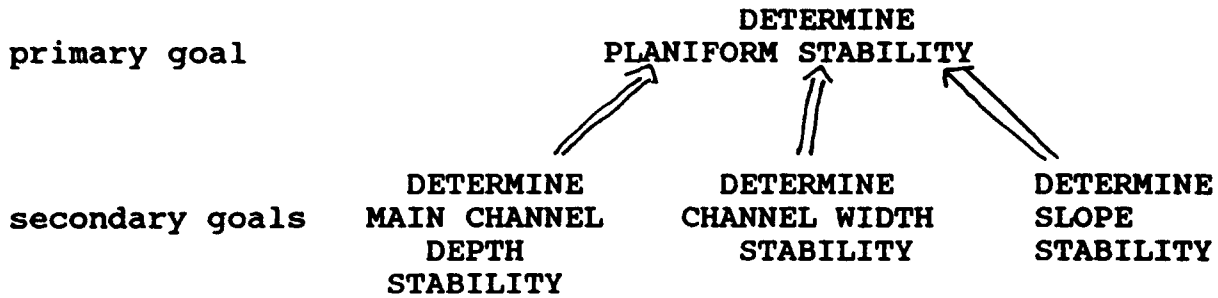
Is it true that:  
A IS true

If the user answers yes, the inference mechanism is able to reach the conclusion that "D IS true" based on the law of hypothetical syllogism<sup>9</sup>.

---

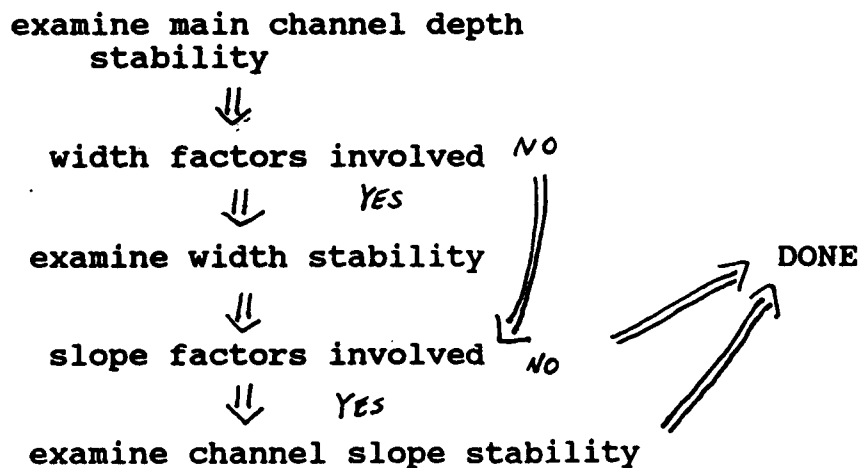
<sup>9</sup> A rule of inference that states: IF A, then B. If B, then C. Therefore, If A, then C.

Suppose we try and create a fully backward chaining problem solving strategy to implement our stream bed flow expert system rule base. We need one primary goal which all the other rules work backward toward solving; a series of secondary goals each of which has information needed by the primary goal; and a whole series of secondary factors which contribute to the information required to satisfy the secondary goals.



secondary factors: channel depth, channel slopes, channel width, fines present on the bar surface, type of bar, etc.<sup>10</sup>

In this simplification of our river channel flow system the flow chart for the problem solving strategy would be:



<sup>10</sup> The secondary factors are connected to the secondary goals by sets of control rules which are explained below.

In order to implement this problem solving strategy in backward chaining we would need what is called control rules. For example, we would write

```

RULE 1
IF Channel Depth and Slope Stability known
AND Channel Width Stability known
THEN Planiform Stability known {conclusion 1}

RULE 2
IF Channel Depth Stability known
AND Channel Slope Stability known
THEN Channel Depth and Slope Stability known {conclusion 2}

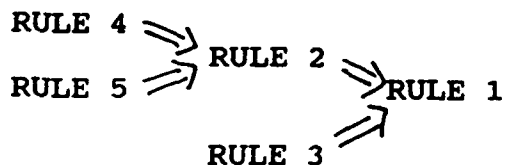
RULE 3
IF Channel Width IS increasing,decreasing,stable
THEN Channel Width Stability known {conclusion 3}

RULE 4
IF Channel Depth IS increasing,decreasing,stable
THEN Channel Depth Stability known {conclusion 4}

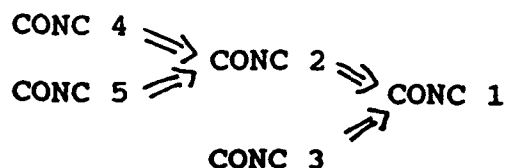
RULE 5
IF Channel Slope IS increasing,decreasing,stable
THEN Channel Slope Stability known {conclusion 5}

```

The goal of the above abbreviated backward chaining system is to arrive at the conclusion "Planiform Stability is known". Since Rule 1 has that conclusion as its consequent, the inferencing mechanism tries to satisfy the two antecedents of Rule 1: "Channel Depth and Slope Stability known" and "Channel Width Stability known". The consequent of Rule 3 matches the latter so the inferencing mechanism tries to satisfy the antecedent of Rule 3: "Width is increasing, decreasing, stable". We would then have a rule dependency map for the full rule set as below:



or a similar graph if we plotted the relationships by rule conclusions instead of rule numbers.



The difficulty with this problem solving strategy is that when we add rules to get the information that is needed to form the conclusions in the higher level rules, the hypotheses for those rules may contain variables which fire other rules that alter previous conclusions. For example, suppose we have a rule that says :

RULE 6

If A diagonal bar is present

THEN Channel Width IS increasing

then we cannot determine what RULE 3 (which involves the Channel Slope in its conclusion) will say until RULE 6 (which involves Channel Width) is evaluated. If the program happens not to consider the rules in the correct order (as it doesn't in this example) then the conclusion reached will not be valid. Thus we see that in the most general situation conclusions reached upon considering beginning rules may have to be reevaluated in light of later conclusions that the rule-based system reaches.

## **How to Create Forward Chaining in a Backward Chaining System**

Although PROLOG goal driven systems are not designed primarily for data driven problems (see Brakto (1986) and Amble (1987) for an explanation of how these systems work) , it is possible to simulate forward chaining in a backward chaining search using global variables<sup>11</sup> and recalling the goal (cycling) after each success. Newer versions (2.5 and later) of the LEVEL5 software, written for IBM compatible personal computers, have significant object oriented features which make forward chaining in this type of system easier. However, LEVEL5 version 1.1 which is available for the MacIntosh and used for this report can also be used for these type of problems.

In cycling the program there must be **only one** top-level goal (which is called "forward chain" in the manual<sup>12</sup>). A global variable (called "step") then allows you to consider different groups of rules on different passes through the rule base. The programmer must then organize the groups of rules so that every group is consider in a fixed sequence of different "steps".

The hierarchy of goal levels shown in the manual where they are listed in outline form such as 1., 1.1, 2., 2.1, only work when you need to make a single pass through the set of rules. First goal selection determines which upper level rule you want the compiler to unify variables<sup>13</sup> on. This works just like when a PROLOG compiler asks you which goal to solve for in its predicate rule base. It does not prioritize the goals and search for all possible solutions. But, the search levels are created by the developer placing what is sometimes called a "salience factor"<sup>14</sup> (or operator precedence factor) on the rule when it is placed on the goal stack<sup>15</sup>. When a new hypothesis is placed on the search

---

<sup>11</sup> "A value established for use when no procedure or binding (a place in memory reserved for a value associated with a symbol) primitive supplies a value." Winston and Horn, op.cit.

<sup>12</sup> LEVEL5 for the Apple Macintosh, User's Guide, by Information Builder's, Inc. 1250 Broadway, New York,N.Y. 10001.

<sup>13</sup> "The process of comparing two pattern expressions to see if they can be made identical by a consistent set of substitutions." Winston and Horn, op.cit.

<sup>14</sup> " A priority number given to a rule. When multiple rules are ready to be satisfied or as is sometimes said, for "firing", they are fired in order of priority. The default salience is zero. Rules with the same salience are fired according to the current conflict resolution strategy." NASA op.cit.

<sup>15</sup> A list of all the goals that the inference mechanism is backward chaining in order to satisfy. The goal at the top of the list is the goal which the compiler is currently searching the



stack<sup>16</sup> (also called an agenda) salience numbers are checked to make the insertion. The search then keeps going until it bottoms out on a lower level goal. Problems suited for this type of program organization are for instance, diagnostic rule bases or classification problems with only one end conclusion. Groups of definitions of objects fall into this category.

One way to order the rule base is to list which rules have variables in their hypotheses, define a partial order on the rules by letting the rules be nodes in a graph and connect two nodes (rules) with an edge if the conclusion of one is used in the hypothesis of another <sup>17</sup>, and then topologically sort the rule base according to this partial order:

#### ALGORITHM TO TOPOLOGICALLY SORT RULES IN AN EXPERT SYSTEM

- 0) For the whole set of nodes of conclusions in the rules:
  - 1) If every conclusion node has a predecessor, then stop. The rule based system has a cycle and is infeasible (that is, a partial order cannot be defined on it).
  - 2) pick a node V which has no predecessor
  - 3) place V on a list of ordered nodes
    - a) if a terminal conclusion node is reached, print out the list of rules used on the way to reach that conclusion.
  - 4) delete all edges leading out from V to other nodes in the network
- 5) Go to step 0).<sup>18</sup>

---

knowledge base of consequents in order to unify variables on.

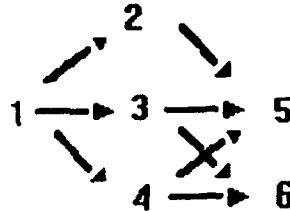
<sup>16</sup> "A list of all rules that are presently ready to be satisfied. It is sorted by salience values and the conflict resolution strategy. The rule at the top of the search stack or agenda is the next rule that will fire.", NASA, op. cit.

<sup>17</sup> Recall the definition from mathematics that a partial order is a relation  $rel(x,y)$  between objects in a set that satisfies the reflexive ( $rel(x,x)$  is always true), and transitive conditions ( $rel(x,y)$  and  $rel(y,z)$  true implies that  $rel(x,z)$  true). If the relation also satisfies the symmetric condition ( $rel(x,y)$  true implies  $rel(y,x)$  true) then it is called an equivalence relation.

<sup>18</sup> A further discussion of the way this algorithm works in the case of any partial order and how to write the pseudo code for a simple version of it is given in the books: Fundamentals of Data Structures by E. Horowitz and S. Sahni, Computer Science

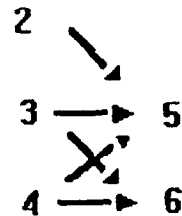
**Example:**

a) initial network:



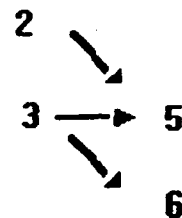
b) node visited - 1

remaining network:



c) node visited - 4

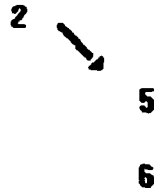
remaining network:



Press, Rockville, MD, 1982., and Algorithms + Data Structures = Programs by Niklaus Wirth, Prentice Hall, 1976. A C source code implementation of it along with a further discussion is included in Appendix II.

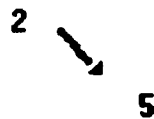
**d) node visited 3**

**remaining network:**



**e) node visited 6**

**remaining network:**



**at this point a terminal conclusion is reached and the number of levels of rules needed (3) is printed out**

**f) node visited 2**

**remaining network:**

**5**

**g) node visited 5**

An example of how the above algorithm works is illustrated in figures 1 and 2 above.

First a successor list for each rule conclusion node is created:

successor list for vertex	1	[ 2 3 4 ]
successor list for vertex	2	[ 5 ]
successor list for vertex	3	[ 5 6 ]
successor list for vertex	4	[ 6 5 ]
successor list for vertex	5	[ ]
successor list for vertex	6	[ ]

Then the algorithm produces a topological ordering of vertices as shown in the figures and as listed below:

1    4    3    6  
terminal conclusion reached  
3 levels of rules required

2    5  
terminal conclusion reached

### 3 levels of rules required

With the above topological ordering when rule 6 is considered, the information for either rule 3 and 4 is then required. And each of these rules will require the information from rule 1. This path of rules then forms at least two forward chaining "cycles" or "steps". We do not know beforehand whether rule 3 or rule 4 will provide the way to satisfy rule 6, hence a separate step is required to recycle through the rules to cover all possible cases (see the next section for the details of how this is implemented in the code). To better organize things for future additions to the rule base, it is prudent to add another step for rule 1 and thus use three steps for this path of rules. For step 4 in this example we consider rule 5. This rule then requires the information from rule 2 and rule 1 in that order. Rule 1 has already been considered in step 1. With this path all rules have been considered. Therefore two more steps of recycling through the rules are required to consider the whole rule base. These two steps will then insure that all the information necessary to reach any possible conclusion has been entered.

If, in entering the input information, we change the order in which the nodes coming out of a given vertex are ordered the program gives a different output. This is a result of the fact that for a given set of order relations there may be many different ways of defining a partial order on them. Consider what happens in the above algorithm if we change the input order:

```
successor list for vertex 1 [ 4 3 2]
successor list for vertex 2 [ 5]
successor list for vertex 3 [ 6 5]
successor list for vertex 4 [ 6 5]
successor list for vertex 5 []
successor list for vertex 6 []
```

the algorithm will then produce the following topological ordering of vertices:

```
1 2 3 4 5
terminal conclusion reached
3 levels of rules required

6
terminal conclusion reached
3 levels of rules required
```

However, upon examining this output, it can be seen that the number of paths, "cycles", or "steps", required to enter all the information into the classification system is the same in the two cases. Also, the maximum depth or number of levels of backward

reasoning for all cycles will be the same for both cases. The next section will explain the coding procedures for implementing these forward chaining cycles or steps which are determined by the topological order.

## **Procedures to Use in Creating the Knowledge Base in a Classification Type Expert System**

In order to create a knowledge base there is an organized procedure that one can follow:

a. Establish the facts by:

(1) collecting all the relevant facts and information

(2) divide the objects in the facts into different categories

(3) outline or catalog the complete set of facts according to these categories.

(4) write down all the rules (involving forward chaining) relating these categories.

(5) write down a decision tree for what the expert system is trying to analyze such as that shown in the previous paragraph.

(6) determine the one goal which the expert system is trying to satisfy.

(7) write down all the backward chaining rules which help to satisfy that one goal.

(8) relate the forward chaining and backward chaining rules in order to have the expert system perform its task in one program run.

i) write down a complete table of all the variables and conclusions involved in the knowledge base.

ii) topologically sort the rules based on the order the conclusion occur in.

iii) plot all the paths in the knowledge base in which variables can be instantiated and conclusions reached (this is done in paragraph 20 in this report).

iv) use global variables to group the instantiation of the variables and predecessor conclusions involved in the hypotheses.

For an example that explains the forward chaining procedure consider the following abbreviated example using the river

(5)

If we add a rule or topic for obtaining the initial information we need to start the forward chaining, and a final rule or topic to generate the report and write it to a file we can now write down an outline or decision tree for what the expert system is trying to analyze.

1. Introduction	
1.1 River's name entered	RULE 1
2. Draw conclusions	
2.1 Bar composition determined	RULE 2    RULE 3
2.2 Channel depth determined	RULE 4
3.	
3.1 Report filed	RULE 5

(6) In this case the one goal that the expert system is trying to satisfy is to generate the final report (which by the way should contain all the conclusions the forward chaining has generated)

If we call this one goal "forward chain"

We can now write down a flow diagram for the program to reach all the conclusions we want:

step 1

river's name

step 2

bar composition

step 3

channel depth

step 4

file a report containing all the conclusions  
reached

(7) We can write this rule to finish the forward chaining  
as:

```
RULE 5
IF previous steps complete
AND FILE results file footer
THEN stop
```

The problem at this point is that there hasn't been a condition

channel flow information<sup>19</sup>:

(1). We have various attributes that describe the river channel flow geometry: among these are bar composition framework gravel, censored gravel, filled gravel, or matrix gravel, channel depth(a numerical value). Once we know these attributes we have a list of rules relating them from which we can draw inferences.

(2) We define a data structure:

ATTRIBUTE The bar composition  
AND channel depth

We will also need a string variable which hold's the river's name to write on the final report:

STRING The river's name

(3) We can now organize this information in the following outline

1. River's name entered
2. Bar composition determined
3. Channel depth determined

(4) we can now write down the forward chaining rules involving these attributes and variables:

RULE 1  
IF The river's name <>""  
AND FILE the results  
THEN River's name entered

RULE 2  
IF Fines are present on the bar surface  
THEN The bar composition IS matrix gravel  
AND Bar composition determined

RULE 3  
IF NOT Fines are present on the bar surface  
AND Grains on the bar surface are interlocked with  
voids  
THEN The bar composition IS Framework Gravel  
AND Bar composition determined

RULE 4  
IF Channel Depth > 0  
THEN Channel Depth determined

---

<sup>19</sup> The steps in the example are ordered according to the letters in the above procedures.



determined which in all cases signals that the forward chaining has ended and tells RULE 5 when to fire.

(8) In a LEVEL5 program it is possible to have series of nested goals for which we use the goal select feature to choose which ones the system backward chains to satisfy. But, in the above example the LEVEL5 compiler will choose either

1.	or	2.	or	3.
1.1		then		3.1.
		2.1 or 2.2.		

But, it will not be able to search for all combinations of solutions to the whole list on the same program run. For once a bottom level goal is reached and there are no more rules to check on the rule search stack (agenda) for the satisfaction of this goal the program stops. That is, suppose the user chooses 1. and 1.1 using goal select and suppose RULE 1 is of the form:

IF .....

THEN Rivers name entered.      We have:

goal stack

1.1 River's name entered  
(satisfied)

rule search stack

RULE 2  
RULE 3  
RULE 4  
(no further matches  
possible)

Then, none of the remaining rules satisfies the goal and the program stops.

The way 'cycling' works is through the use of a single goal along with global variable values. In the above example we would have only one goal:

goals

1. forward chain

In rewriting RULE 1 we substitute 'forward chain' for the previous goal and add a global variable 'step'.

RULE 1

IF The river's name <>"  
and FILE results file header  
and step:=1  
THEN River's name entered  
and step:=2  
and CYCLE

global variable  
step

RULE 2

IF Fines are present on the bar surface

```

and step:=2
THEN The bar composition IS matrix gravel
and Bar composition determined
and step:=3
and CYCLE

```

```

RULE 3
IF NOT Fines are present on the bar surface
and Grains on the bar surface are interlocked and no voids
and step:=2
THEN The bar composition IS Framework Gravel
and step:=3
and Bar composition determined
and CYCLE

```

```

RULE 4
IF Channel Depth > 0
and step:=3
THEN Channel depth determined
and forward chain
and step:=4
and CYCLE

```

```

RULE 5
if step:=4
and FILE the results
then stop
built-in function stop

```

In this situation after the goal forward chain has been reached for the first time the global variable step is changed from 1 to 2. Then, there are more rules left in the search stack (agenda) which can satisfy the same goal, so the program continues to search for solutions:

goal stack	search stack
1. forward chain (satisfied	RULE 2
	RULE 3
	RULE 4
global variable	(further matches possible)
step	
built-in function	
stop	

The way the program finally stops is by means of using a built-in function. The phrase 'built-in' means that the program will execute

its meaning as soon as it is scanned<sup>20</sup> and before the whole sentence is parsed<sup>21</sup>.

note: In this example because of the small number of rules it is not necessary to perform substeps i), ii), and iii in part (8) of the procedure. The next example shows how using the basic techniques in parts (1) through (7) along with the substeps in part (8) it is possible to organize a large scale system.

We now consider the group of all the rules in our river channel flow expert system in which the variables and conclusions are to be chained together:

For these purposes a variable is defined to be a statement which appears in the hypothesis section of a rule and about which it is to be determined whether it is true or false or instantiated to some object or attribute<sup>22</sup>.

We now consider the group of all the rules in our river channel flow expert system in which the variables and conclusions are to be chained together:

---

<sup>20</sup> The scanner is the part of the compiler that analyses characters of the program's text for identification of known words, variables, functions, and procedures.

<sup>21</sup> The parser is that part of a compiler that analyses complete sentences of the knowledge base (rules, facts, etc.) and determines their total meaning.

<sup>22</sup> A variable is instantiated when there is some object or attribute which it is equal to.

Table 2

Variable List	
Number	Variable
V1	Fines are present in the reach
V2	Fines are present on the bar surface
V3	Large clasts are in direct contact
V4	Vegetation is present on the bar surface
V5	Bar stability is known
V6	The bank is eroding
V7	The bar is growing
V8	Channel Depth is known
V9	Slopes are known
V10	The type of bar has been identified
V11	Active channel width is known
V12	Fresh scour can be seen on the outside bank
V13	Fresh scour can be seen on the bar
V14	The reach is a river bend
V15	Grains on the bar surface are interlocked
V16	Fresh deposits can be seen on the bar
V17	Fines are present in the reach
V18	The type of bar present
V19	The bar composition
V20	The active channel width
V21	The bar is migrating
V22	Diffuse gravel sheets exist
V23	Imbrication is present
V24	bar deposition will occur

Table 3

Conclusion List		
Number	Conclusion	Hypotheses <sup>23</sup>
C1	The bar composition IS----	V2,V3,V15
C2	The type of bar present IS ----	
C3	The bar is providing transported sediment.	C15
C4	The bar is eroding faster than the bank.	V13,C15
C5	The active channel width IS ---	C11,C15,C7
C6	The main channel depth IS ---	C11,C15,C7,V23,C1,C17,V24
C7	The bar is growing	C14,C2,V16
C8	Width conclusion IS ----	V12,C2,V14,C4,C17,C15
C9	The active channel width will be ---	C2,V12
C10	The bar may be migrating	V2,C2,V17,C15,V4
C11	The bank is eroding	V12
C12	The bank is providing transported sediment	C11
C14	The bar is migrating	C2,V2,V17,V4
C15	The bar is eroding	C2
C16	Deposition is occurring on the inside of the bend	C2,C11,V14
C17	Sediment transport is -----	C4,C12
C20	Bar deposition will occur	V22
C23	The slope IS -----	C5

<sup>23</sup> Variables which appear in a rules hypothesis are numbered V1,V2, etc. according to the number appearing in the variable table, conclusion are numbered C1,C2,etc analogously. The information for conclusions which do not have variables or other conclusions listed in this column is entered interactively.

C24	The channel stability is -----	C23
C26	Meander pattern IS -----	C23

Table 4

Rule List <sup>24</sup>		
Number	Rule Name	Conclusions
R1	Identify bar	C2
R2	Forcing a few more conclusions	C7,C11
R3,R4,R5 ,R6	Composition	C2
R7	Bank Scour indicates erosion	C11
R8	Scour indicates erosion	C15
R9	Bar erosion provides sediment	C12
R10	Bank erosion provides sediment	C12
R11	Bar eroding before bank	C14
R12	Stable channel with respect to width	C5
R13	Bar erosion implies increasing width	C5
R14	Wider channel implies less depth and energy	C6
R15	Bank erosion implies increasing width	C5
R16	Growing bar	C7
R17	Narrowing channel	C5,C6
R18,R19, R20,R21	Eroding bar indicates not interlocked Relative shear force Bank scour indicates migration to outside No width conclusions can be reached	C8

<sup>24</sup> This is an abbreviated list of the total number of rules in the knowledge base. In order to simplify the table, some groups of rules which have no other rule predecessors have been grouped together as single rules and rules which only perform functions such as displaying text and pictures or writing the report have been omitted.

R22	Deposition forming classical bar	C8,C16
R23	Bar scour indicates greater channel width	C9
R24	Eroding bar may be migrating	C10
R25	Fines indicate that bar is not migrating	C10,C14
R26	Absence of fines indicates migrating bar	C10
R27	Vegetation indicates stable bar	C10
R28	Sediment transport increasing	C17
R29	Diffuse gravel sheets exist	C20,C23
R30	Decreasing channel width increases slope	C23,C17
R31	Increasing channel width decreases slope	C23,C17
R32	Increasing slope decreases meander pattern	C26,C21,C22
R33	Decreasing slope increases meander pattern	C21,C22,C26,C17
R34	Stable channel with respect to depth	C6
R35	Imbrication indicates bed stability	C6
R36,R37	Framework or matrix gravel affects depth of main channel	C6
R38,R39	Affect of sediment transport on depth	C6
R40	Bar erosion implies less depth	C6
R41	Narrowing channel increases depth	C6



A table (or edge adjacency list<sup>23</sup>) can now be made with the hypothesis conclusion vertex numbers from the right column of the conclusion list table above on the left. On the right is put the conclusion list vertex numbers from the left hand column of the conclusion list table above. A summary of the information from the core rules in the expert system that was entered into the the topological sort program in order to organize its knowledge base is given below:

C15 => C3	C2 => C9
C15 => C4	C2 => C10
C11 => C5	C15 => C10
C15 => C5	C11 => C12
C7 => C5	C2 => C14
C11 => C6	C2 => C15
C15 => C6	C2 => C16
C7 => C6	C2 => C11
C1 => C6	C4 => C17
C17 => C6	C12 => C17
C14 => C7	C5 => C23
C2 => C7	C23 => C24
C2 => C8	C23 => C26
C4 => C8	
C17 => C8	
C15 => C8	

---

<sup>23</sup> An edge adjacency list is a list containing all the end nodes from a given start node (or vertex) in the rule bases' knowledge tree. This is a general term to designate a data structure used in network search algorithms. For the example given the context being discussed the first letters (C) from each node have been removed and a list made of all connections, indexed by the starting node.

The topological sort program then orders the rule numbers according to the priorities in their data dependencies.

After the information from the conclusion list table is entered the program outputs this edge adjacency list:

successor list for rule hypotheses

```

vertex 1 [ 6 ]
vertex 2 [ 7 8 9 10 14 15 6 11 ]
vertex 3 [ ]
vertex 4 [ 8 17 ]
vertex 5 [ 23 ]
vertex 6 [ ]
vertex 7 [ 5 6 ]
vertex 8 [ ]
vertex 9 [ ]
vertex 10 [ ]
vertex 11 [ 5 6 12 ]
vertex 12 [ 17 ]
vertex 13 [ ]
vertex 14 [ 7 ]
vertex 15 [ 3 4 5 6 8 10 ]
vertex 16 [ ]
vertex 17 [ 6 8 ]
vertex 18 [ ]
vertex 19 [ ]
vertex 20 [ ]
vertex 21 [ ]
vertex 22 [ ]
vertex 23 [ 24 26 ]
vertex 24 [ ]
vertex 25 [ ]
vertex 26 [ ]

```

Finally, after the program has established the partial order of the conclusions by their data dependencies, the rest of the rule edges in the edge adjacency list at the beginning of the section can be added to complete the knowledge tree:



## BIBLIOGRAPHY

Amble, T., Logic Programming and Knowledge Engineering, Addison Wesley Publishing Co., 1987.

Ashmore, Peter E. (1985) Process and Form in Gravel Braided Streams: Laboratory Modelling and Field Observations. Thesis for the University of Alberta.

Barker, D., Developing Business Expert Systems with LEVEL5, MacMillan Publishing Co., New York, N.Y., 1988.

Brakto, Ivan, PROLOG, Programming for Artificial Intelligence, Addison Wesley Publishing Co., 1986.

Carson, M.A., 1984 Observations on the meandering-braided river transition, the Canterbury Plains, New Zealand. New Zealand Geographer 40: 12-17, 89-99

Charnak, E, Riesbeck, C, McDermott, D. Meehan, J., Artificial Intelligence Programming, 2nd ed., Lawrence Erlbaum Associates, Hillsdale, N.J. ,1987.

Giarratono, Joseph C., CLIPS User's Guide, NASA Lyndon B. Johnson Space Center, Information Systems Directorate, Software Technology Branch, 1991.

Harrell, A. W., Organizing the Knowledge Base of A Classification Expert System, Technical Report GL-93- , Waterways Experiment Station, Vicksburg, MS, 1993, to appear.

Hey R.D., Bathurst, J.C. and Thorne, C.R. (Editors), (1982) Gravel-bed Rivers. Wiley: 867p.

Horowitz, E. and Sahni, S., Fundamentals of Data Structures, Computer Science Press, Rockville, MD, 1982.

Information Builder's, Inc. LEVEL5 for the Apple Macintosh, User's Guide, 1250 Broadway, New York, N.Y. 10001.

Laronne, J.B. and Maddock, T., 1953 The hydraulic geometry of stream channels and some physiographic implications. Professional Paper, 252, United States Geological Survey: 57 p.

Lane, E. W., 1955 Design of stable channels. Transactions, American Society of Civil Engineers 120: 1234-1260.

Laronne, J. B. and Carson, M. A. (1976). Interrelationships between bed morphology and bed-material transport for a small,

gravel-bed channel. *Sedimentology*, 23, pp. 67-85.

Leopold, L.B. and Maddock, T. Jr. (1953) The Hydraulic geometry of stream channels and some physiographic implications, Professional Paper. United States Geological Survey, 252, pp. 1-57.

Lisle, Thomas E. 1986 Stabilization of a gravel channel by large streamside obstructions and bedrock bends, Jacoby creek, northwestern California. *Geological Society of America Bulletin*. Vol. 97, pp 999-1011.

Lewin, I., 1976. Initiation of bedforms and meanders in coarse grained sediment. *Bulletin, Geological Society of America* 87: 281-285.

Mosely, M.P., 1976. An experimental study of channel confluences. *Journal of Geology* 84:535-562.

Mosely, M.P., 1982. Analysis of the effect of changing discharge on channel morphology and instream uses in a braided river, Ohau River, New Zealand. *Water Resources Research* 128: 800-812.

Pederson, Ken, *Expert Systems Programming, Practical Techniques for Rule Based Systems*, John Wiley and Sons, New York, N.Y. 1989.

Winston, P.H., and Horn, B., *LISP 3rd ed.*, Addison Wesley Publishing Co., 1989

Wirth, N, *Algorithms + Data Structures = Programs*, Prentice Hall, 1976.

#### ACKNOWLEDGEMENT

The tests described and the resulting data presented herein, unless otherwise noted were obtained from research conducted under the MILITARY RESEARCH DEVELOPMENT TEST AND EVALUATION PROGRAM of the United States Army Corps of Engineers by the U.S. Army Engineer Waterways Experiment Station. Permission was granted by the Chief of Engineers to publish this information.

# THEORETICAL ALGORITHMS FOR SOLVING THE ARMY STATIONING PROBLEM

Janet Hurst Spoonamore, Ph.D.

U.S. Army Construction Engineering Research Laboratory

P.O. Box 9005 Champaign, IL 61820-9005

217-373-7267 email: spoonam@csrd.uiuc.edu

**Abstract.** This research addresses algorithmic approaches for solving the Army stationing problem. The problem is formulated as an assignment problem with the objective function being a minimization problem. The specific assignment problem has piece-wise linear additive separable server cost functions, which are continuous everywhere except at zero, the point of discontinuity for the  $\{0, 1\}$  assignment condition. Continuous relaxation of the  $\{0, 1\}$  constraints yields a linear programming problem. Solving the dual of the linear programming problem yields the complementarity conditions for a primal solution, a system of linear inequalities and equalities. Adding equations to this system to enforce a  $\{0, 1\}$  solution in the relaxed solution set yields an augmented system, not necessarily linear. Methods to solve this system, a system of linear inequalities and non-linear equations, in a least square sense are developed, extending Han's method for solving linear systems of inequalities. Generalizations of these methods to solve general systems of inequalities in a least square sense are developed. A sample problem is shown.

*Key words:* Assignment problem, least square problem, stability, global convergence, local convergence, almost everywhere differentiable, Clarke subdifferential.

## 1. Formulating the Stationing Problem.

Presently, the military is transitioning from the Cold War era to the new "Power Projection Platforms" to meet the global challenges of the 21st century. Major stationing changes, realignments and force structure reductions are being planned and executed. At the same time, increasing fiscal constraints require attention to cost efficient operations. It is important to consider the least cost options of basing units among all possible installations. By formulating an assignment problem, which is based on cost functions of different unit types at the various installations, one could determine the least cost stationing alternative among all possible alternatives.

Let  $i \in I = \{1, 2, \dots, m\}$  represent military units to be assigned to installations  $j \in J = \{1, 2, 3, \dots, n\}$ . Let  $1^m \in \mathbb{R}^m$  represent the unit demand vector of the  $m$  military units. For each  $j \in J$ ,  $i \in I$ , let  $y_{i,j}$  be the relative amount of demand  $i$ , assigned to installation  $j$ , i.e.,  $y \in \mathbb{R}^{mn}$ . The variable  $y$  is constrained so that  $0 \leq y_{i,j}$ , for each  $j \in J$ ,  $i \in I$ . Further, the variable  $y$  must meet the demand for each  $i$ . That is, for all  $i \in I$ ,

$$\sum_{j \in J} y_{i,j} = 1.$$

Let positive linear cost coefficients,  $c_{i,j}$ , and positive minimum cost constants,  $\phi_j$ , be associated with each installation  $j$  and military unit  $i$ . That is, assume for all  $i \in I$  and  $j \in J$ , that  $c_{i,j} > 0$  and  $\phi_j > 0$ . To simplify notation, interpret the vector  $y_{*,j}$  to be the elements of  $y \in \mathbb{R}^{mn}$ , associated with installation  $j$ . The cost function for each installation  $j$  is:

$$\begin{cases} \max(\phi_j, \langle c_{*,j}, y_{*,j} \rangle) & \text{if any } y_{i,j} > 0, \\ 0 & \text{if } y_{i,j} = 0, \forall i \in I. \end{cases}$$

This installation cost function represents a flat minimum charge with linear rates beyond the minimum.

This assignment problem belongs to a family of problems, called the uncapacitated facility location problem. It is a generalization of the problem described in Conn and Cornuéjols[1990]. See Spoonamore[1992] where the relaxation of the problem is expressed and the relaxed dual is formed. Conn and Cornuéjols[1990] show solving this problem by solving the relaxed dual where they define gradients for determining search directions. The relaxed problem, rather than being solved using the usual linear programming methods, can be solved by expressing the problem as a system of linear equalities and inequalities and solved using the method developed by Han[1982].

Sample Problem:

In order to illustrate the primal and dual relationship, the following simple problem is shown. Let  $m = 4$ ,  $n = 3$ ; let  $\phi = (26, 30, 30)$ ; let

$$c = \begin{pmatrix} 20 & 25 & 25 \\ 6 & 5 & 5 \\ 40 & 40 & 44 \\ 30 & 30 & 22 \end{pmatrix}.$$

The optimal value is  $87\frac{3}{11}$ , achieved at, for example,  $t = (1, 0, 1)$  and  $w = (1, 0, 1)$  and:

$$u = \begin{pmatrix} 20 \\ 5 \\ 40 \\ 22\frac{3}{11} \end{pmatrix}, \quad v = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{5}{11} \\ 0 & 0 & 0 \\ 0 & 0 & 2\frac{3}{11} \end{pmatrix}, \quad y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ \frac{41}{44} & 0 & \frac{3}{44} \\ 0 & 0 & 1 \end{pmatrix}.$$

$s_1 = s_2 = 0$ , and  $s_3 = \frac{1}{11}$ . Thus,  $r_1 = r_2 = 1$ , and  $r_3 = \frac{10}{11}$ . We show that this solution satisfies the optimality conditions (2.3.1).

First, consider  $j = 2$ , where  $y_{i,2} = 0$  for all  $i$ , satisfying  $I_2^0 = \{1, 4\}$ , and  $I_2^w = \emptyset$ . Further,  $w_2 = 0$  is consistent, since  $J^\geq = \{1, 2\}$  and

$$\langle c_{*,2}, y_{*,2} \rangle = 0 \geq w_2 \phi_2 = 0 * 30 = 0.$$

Second, consider  $j = 3$ .  $I_3^0 = \{1\}$  and  $I_3^w = \{2, 4\}$ . It must be that  $r_3 = \frac{10}{11} \in (0, 1)$  and  $J^= = \{3\}$ . It must be that  $y_{4,3} = 1$ , which implies that  $w_3 = 1 = y_{2,3}$ . To satisfy,

$$\langle c_{*,3}, y_{*,3} \rangle = w_3 \phi_3 = 30,$$

then  $y_{3,3} = \frac{3}{44}$ .

Lastly, consider  $j = 1$ , where  $I_1^0 = \{2, 4\}$  and  $I_1^w = \emptyset$ . Since  $1 \in J^\geq$ , then

$$\langle c_{*,1}, y_{*,1} \rangle = 57\frac{3}{11} > w_1 \phi_1 = 26.$$

$$y_{1,1} = 1 \leq w_1 = 1; \quad y_{1,3} = \frac{41}{44} \leq w_1.$$

For  $i = 3$ ,  $\sum_j y_{3,j} = 1$ . Further, for all  $i$ ,  $\sum_j y_{i,j} = 1$ .

The full set of solutions of this problem include:  $y_{i,j} = 0$ , except as follows:

$$y_{1,1} = w_1 = 1, \quad y_{2,3} = y_{4,3} = w_3 = 1, \quad y_{3,3} = \frac{3}{44},$$

and  $w_2, y_{3,2}$ , and  $y_{3,1}$  satisfy:

$$w_2 \leq 1,$$

$$y_{3,1} \geq \frac{6}{40},$$

$$y_{3,2} = \frac{41}{44} - y_{3,1},$$

$$0 \leq y_{3,2} \leq w_2 \leq \frac{4}{3} y_{3,2}.$$



## 2. Least Square Formulation for Solving Systems of Inequalities.

By expressing the solution to a linear programming problem as the solution to a system of inequalities and equalities, one can take advantage of the work of Han[1982], Mangasarian[1981]. For simplicity, we only consider the system of inequalities. Firstly, consider linear systems.

2.1 Statement of Problem: Consider a linear transformation,

$$A : \mathbb{R}^n \mapsto \mathbb{R}^m, \quad b \in \mathbb{R}^m, \quad x \in \mathbb{R}^n.$$

One wishes to solve for  $x \in \mathbb{R}^n$ , such that

$$Ax \leq b. \quad (2.1)$$

The least square formulation of (2.1) yields

$$\min_x f(x) := \frac{1}{2} \langle (Ax - b)_+, (Ax - b)_+ \rangle, \quad (2.2)$$

where  $(Ax - b)_+$  denotes the projection of  $Ax - b$  onto  $\mathbb{R}_+^m$ .

2.2 Numerical Stability: Robinson[1975,1976] shows the numerical stability properties of solution sets of systems of inequalities. Let  $\Omega := \{x | Ax \leq b\}$ . Define  $\Omega$  as *stable* if for all  $x_0 \in \Omega$ , there exists  $\beta, \delta$  such that for any  $A' : \mathbb{R}^n \mapsto \mathbb{R}^m, b' \in \mathbb{R}^m$  satisfying

$$\|A - A'\| + \|b - b'\| \leq \delta,$$

then

$$\text{dist}(x_0, \Omega') \leq \beta \rho(x_0),$$

where

$$\Omega' := \{x | A'x \leq b'\}, \quad \rho(x) := \inf_{k \in \mathbb{R}_+^m} \|b' - A'x - k\|.$$

Define the system  $Ax \leq b$  as *regular* if there is an  $x \in \mathbb{R}^n$  such that  $Ax < b$ .

Robinson shows that  $\Omega$  being stable is equivalent to the system  $Ax \leq b$  being regular. The optimality conditions for a solution  $x, z$  which solves

$$\begin{aligned} \inf_{x,z} \quad & \frac{1}{2} \langle z, z \rangle \\ \text{subject to:} \quad & Ax - b \leq z \end{aligned}$$

are

$$\begin{aligned} A^T z &= 0, \\ z &\geq 0, \\ Ax - b - z &\leq 0, \\ \langle z, Ax - b - z \rangle &= 0. \end{aligned} \quad (2.3)$$

Thus,  $\Omega := \{(x, z) | z \geq 0, Ax - b - z \leq 0\}$  is stable since the system  $z \geq 0, Ax - b - z \leq 0$  is regular.

### 3. Local Convergence Properties of Taylor Series based on Generalized Differential Constructs.

In this section, we show that, by defining generalized differential constructs for the least square objective function and the projection, one can take advantage of strong approximating properties in the Taylor series expansion based on these constructs. Most importantly, we first show that these constructs are well defined. Note that  $(Ax - b)_+$  and  $\nabla f(x)$  are both Lipschitz continuous functions throughout all  $\mathbb{R}^n$ . Thus, the Clarke generalized Jacobian,  $\partial(Ax - b)_+$ , and generalized Hessian,  $\partial\nabla f(x)$ , both exist and are well-defined. Recall that  $J(x) \in \partial(Ax - b)_+$  means:

$$(J(x))_i^j = \begin{cases} a_i^j & \text{if } \langle a_i, x \rangle - b_i > 0, \\ \lambda_i a_i^j & \text{if } \langle a_i, x \rangle - b_i = 0, \text{ where } \lambda_i \in [0, 1], \\ 0 & \text{if } \langle a_i, x \rangle - b_i < 0. \end{cases}$$

Similarly,  $H(x) \in \partial\nabla f(x)$  is determined by choice of  $J(x) \in \partial(Ax - b)_+$  where  $H(x) = J^T(x)J(x)$ . See Spoonamore[1992] where the following proposition is proven.

**Proposition: Existence of an Identification Neighborhood.**

Let  $x_*$  be given and let

$$I^0(x_*) := \{i \mid \langle a_i, x_* \rangle = b_i\},$$

$$I^+(x_*) := \{i \mid \langle a_i, x_* \rangle > b_i\},$$

$$I^-(x_*) := \{i \mid \langle a_i, x_* \rangle < b_i\}.$$

Then there exists a neighborhood  $N(x_*, r)$  of  $x_*$ , such that for  $x \in N(x_*, r)$ :

$$I^0(x) \subset I^0(x_*),$$

$$I^+(x_*) = I^+(x) \setminus I^0(x_*), \quad (3.1)$$

$$I^-(x_*) = I^-(x) \setminus I^0(x_*).$$

This neighborhood is referred to as the *identification neighborhood* of the point  $x_*$  with respect to the system of inequalities  $Ax \leq b$ . The approximating properties of the generalized Jacobian and the generalized Hessian are shown in the following propositions which are developed in Spoonamore[1992].

**Proposition. Perfect Approximating Property of the Generalized Jacobian.**

Let  $x_*$  be given and let  $N(x_*, r)$  be a neighborhood which satisfies (3.1) in the proposition, above. Let  $x \in N(x_*, r)$  and let  $p := x_* - x$ , then for any  $J(x) \in \partial(Ax - b)_+$ ,

$$(A(x + p) - b)_+ - (Ax - b)_+ - J(x)p = 0.$$

**Proposition. Perfect Approximating Property of the Generalized Hessian.**

Let  $x_*$  be given and let  $N(x_*, r)$  be an identification neighborhood as above. Let  $f(x) := \frac{1}{2}((Ax - b)_+)^T(Ax - b)_+$ . Let  $x \in N(x_*, r)$  and let  $p := x_* - x$ . Then for any  $J(x) \in \partial(Ax - b)_+$ ,

$$f(x + p) - f(x) - \langle p, \nabla f(x) \rangle - \frac{1}{2} \langle p, J^T(x)J(x)p \rangle = 0.$$

#### 4. Conclusion

In Spoonamore[1992], these properties are used to develop optimization algorithms which parallel the existing algorithms for the differentiable case. The Army stationing problem can be solved by formulating the problem as an assignment problem having a piece-wise linear objective cost function. The relaxation of the problem into a linear programming problem allows solution using several methods including the method based on solving systems of inequalities.

#### Bibliography

- F. H. Clarke, *Optimization and Nonsmooth Analysis*, Les Publications CRM, Université de Montréal, 1989.
- A. R. Conn and G. Cornuéjols, "A Projection Method for the Uncapacitated Facility Location Problem," *Mathematical Programming* 46 (1990), 272-298.
- S. P. Han, "Least Square Solution of Linear Inequalities", Unpublished Manuscript, University of Wisconsin at Madison, 1982.
- O. L. Mangasarian, "Iterative Solution of Linear Programs," *SIAM Journal on Numerical Analysis* 18 (1981), 606-614.
- S. M. Robinson, "Stability Theory of Systems of Inequalities. Part II: Differentiable Nonlinear Systems," *SIAM Journal on Numerical Analysis* 13 (1976), 497-513.
- S. M. Robinson, "Newton's Method for a Class of Nonsmooth Functions," *SIAM Journal on Numerical Analysis* 13 (1990) .
- J. H. Spoonamore, *Least Square Methods for Solving Systems of Inequalities with Application to an Assignment Problem*, Ph.D. Dissertation, University of Illinois, Urbana-Champaign, Urbana, IL, November 1992.

# **Severely Constrained Allocation of a Bounded Number of Transceivers**

**T. Cronin**  
**Intelligence Electronic Warfare Directorate**  
**Warrenton VA 22186-5100**

One of numerous Army applications for asset management is to move a network of transceivers belonging to friendly forces into position to monitor the communications of a similar network of transceivers managed by an opponent. Many previous attempts to solve the bounded resource allocation problem have been excursions into the realm of unconstrained optimization, whereas other attempts have framed the problem as one of constraint satisfaction. The two approaches are seemingly at odds with each other, since it has been a moot point about whether constraints be utilized explicitly, or incorporated into an objective function. This paper attempts to resolve the conflict by showing that the two approaches are complementary, albeit at opposite ends of the algorithmic complexity spectrum. When no constraints are available to control the placement of  $k$  transceivers, the challenge of the problem is to maximize the disjunction of  $k$  fields of view, and it is shown that the time complexity is exponential in the number of transceivers. When the locations of opposing force transceivers are known, together with the radio frequency (RF) propagation graph, the problem reduces to Knuth's stable marriage theorem, and the resultant complexity is linear in the number of transceivers. In addition to a characterization of the computational complexity of the problem, two other results have emerged from this research: a novel strategy to allocate relays and transceivers to the fringes of invisible areas; and a massively parallel architecture to compute a comprehensive ensemble of field of view bitmaps.

## **Statement of the problem.**

In diverse terrain, when using a network of friendly transmitter/receiver (transceiver) devices to collect data being communicated by a network of transceivers controlled by an adversary, there is a dynamic requirement to spatially configure the network in such a way that friendly transceivers collect maximal information from the adversary without forsaking the ability to communicate among themselves. The control knowledge to guide the optimization process is in large part a function of what one can cooperatively see or hear from various vantage points of the terrain, which formulates a specialized problem in resource allocation. The allocation process may or may not be facilitated by a variety of constraints, depending upon their availability. If few constraints present themselves, then the problem is hard; on the other hand, if constraints are readily available, then the problem is made correspondingly simpler. The flip side of the collection problem, called *jamming*, is relevant when it is desired to deny adversarial communication, rather than to collect it. The overall problem of tasking a bounded number of communications devices to collect and/or jam adversarial communications is called the Intelligence Electronic Warfare (IEW) asset management problem. In the discussion below, the word "resource" or "asset" refers to either a transceiver or a jammer, whether it be ground-based or airborne.

## **The issues of line of sight and field of view.**

*Line of sight* (LOS) is a concept which distinguishes between what parts of a map are visible and invisible along a given line to an observer located at a specific vantage point. *Field of view* (FOV), also called *area coverage* (AC), is the union of all lines of sight radiating outward from a specific vantage point. For either the constrained or unconstrained transceiver placement problem, LOS and

FOV are fundamental operations. Within the Department of Defense, there has been a proliferation of line of sight algorithms, which has recently prompted a standardization study [J1]. There are two types of line of sight algorithms: *optical* and *electronic*. The optical version renders a field of view display based on what is visible to a rangeless optical sensor. Vegetation, among other factors, may cause seasonal variations. The electronic version addresses a harder problem in that it takes into account radio frequency (RF) propagation and associated power loss. It too, is affected by vegetation and by other factors, such as the attenuating effects of a rough vs. smooth earth, troposcatter, diffraction, reflection, soil conductivity, and solar flux. One of the more sophisticated electronic line of sight algorithms is the Terrain Integrated Rough Earth Model (TIREM) [S1].

In Figure 1, a simulated terrain and corresponding topographic map are depicted to illustrate optical line of sight concepts. In the topographic map, light-colored areas are at higher elevation. The terrain consists of hilltops A and B connected by saddle E, which forms a divide between valleys C and D. Vantage point C offers vistas of hills A and B, and saddle E, but does not provide views beyond. The field of view bitmap from point C, denoted  $\lambda_C$ , is characterized by the diagonal line running from the lower left corner of the topographic map to the upper right corner, portrayed at the left of Figure 2. Visible areas are colored white, whereas invisible areas are gray. From C, one cannot see over into valley D, nor from D can one see into valley C. Note that vantage point D's line-of-sight bitmap,  $\lambda_D$ , is the complement of C's. It is important to observe that taken together (the set union  $\lambda_C \cup \lambda_D$ ), the combined vista offered by C and D is as comprehensive as that offered by points A, B, or E separately.

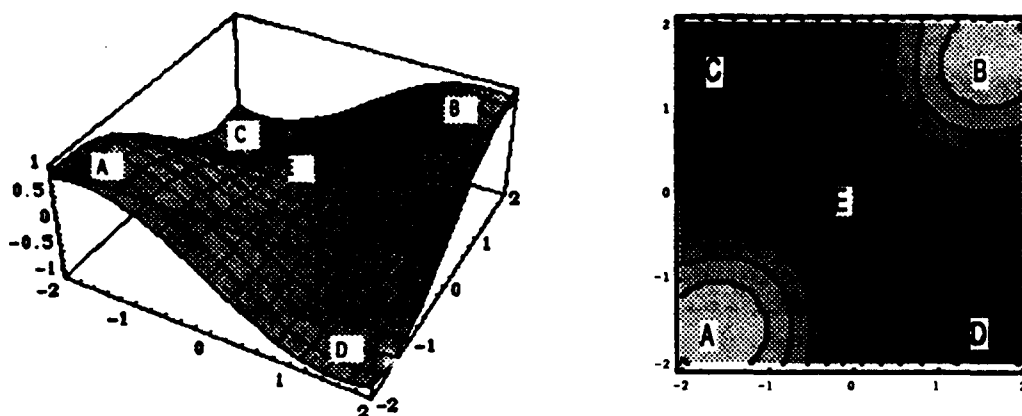


Figure 1. A simulated terrain, and corresponding topographic line map.



Figure 2. Optical field of view bitmaps from valleys C and D respectively (white is visible).

All field of view algorithms currently under development by the Department of Defense utilize Digital Terrain Elevation Data (DTED) as input to generate field of view bitmaps. DTED, produced by the Defense Mapping Agency (DMA), is a multi-megabyte gridded database. A gridded database is a discrete representation of a continuous terrain, where an elevation value is associated with each grid cell of the map of interest. In Figure 3, the topographic map seen earlier has been overlaid with a ten by ten cell grid. Some of the grid cells have been labeled with their respective elevation values to illustrate a gridded database. Two different grid resolutions are currently available from the Defense Mapping Agency: DTED Level 1, representing one hundred meter horizontal spacing, and DTED Level 2, with thirty meter spacing. DTED Level 2 provides a finer resolution product, but the computer memory overhead is more prohibitive.

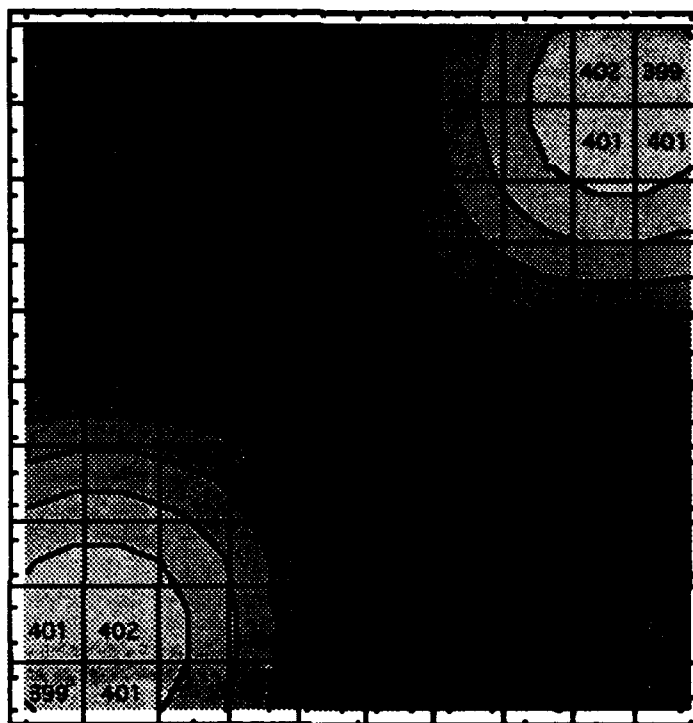


Figure 3. A gridded elevation database.

When constructing a field of view bitmap, any algorithm which accesses each element of the digital grid is said to be an *exact* algorithm. With current technology, the time complexity required to produce an exact optical field of view is  $O[n^3]$ , where  $n$  is the number of sampled elevation values along one edge of the DTED database. If some elements of the digital grid are ignored or bypassed during processing, then the algorithm is said to be approximate. There are a variety of algorithms of time complexity  $O[n^2]$  which produce an approximate field of view; for examples, refer to [B1, R1]. The Broome algorithm, which utilizes a moving horizon technique, was the first to achieve  $O[n^2]$  time complexity. Empirical data indicate that the Ray algorithm produces output matching the exact with high probability [R2]. On a single processor, it is easy to see that the time complexity to produce a set of exact field of view bitmaps for an  $n \times n$  gridded database is  $O[n^5]$ , while the time complexity to produce a set of approximate field of view bitmaps is  $O[n^4]$ .

Electronic line of sight algorithms combine power loss with terrain to produce a set of power contours emanating outward from a sensor placement position. Generally, electronic line of sight is more encompassing than optical line of sight, because electronic signals can be heard in places which cannot

be seen. For the transceiver allocation problem, it is important that two potential transceiver sites lie within the same power contour at a specified signal-to-noise-level. If a transmitter lies beyond the contour, it cannot be heard by another transceiver. Also, if all other transceivers lie outside a power contour corresponding to a specific transceiver location, then the transceiver will not be able to communicate at that power level with neighboring transceivers.

### Global visibility and the unconstrained problem.

For asset management applications, it is useful to quantize field of view by devising a metric which indicates global visibility of the bitmap. If there are  $w$  visible grid cells on an  $n \times n$  bitmap  $\lambda$ , then the *visibility ratio* of the bitmap, denoted  $v_r(\lambda)$ , is defined to be:

$$v_r(\lambda) = w/n^2 \quad [1]$$

The visibility ratio is a metric useful for transceiver allocation during regimes of unconstrained optimization, when fields of view from several vantage points must be cooperatively combined to achieve maximal global visibility. When the allocation process cannot avail itself of the search reduction afforded by constraint satisfaction, then it must resort to an unconstrained optimization scheme. In the unconstrained transceiver placement problem, the following objective function must be maximized, where  $\lambda_i$  is the field of view bitmap corresponding to the  $i^{\text{th}}$  grid cell (where  $i$  is incremented from left to right, then down), and  $k$  is the number of transceivers allocated:

$$v_r \left[ \bigcup_{i=1}^k \lambda_i \right], \text{ where } \lambda_i = \lambda_j, 1 \leq j \leq n^2 \quad [2]$$

**Theorem.** For the unconstrained transceiver allocation problem, the time complexity required to optimally place  $k$  transceivers is  $O[n^{2k}]$ , where  $n$  is the number of grid cells along one edge of the gridded elevation database.

**Proof.** On a gridded database with dimensions  $n \times n$ , the number of ways to place  $k$  transceivers is the combination of  $n^2$  objects grouped  $k$  at a time:

$$\binom{n^2}{k} = \frac{n^2(n^2-1) \cdots (n^2-k+1)}{k(k-1) \cdots 1} = O[n^{2k}] \quad [3]$$

This result is discouraging, since it indicates that as the number of transceivers to be allocated increases, the time required to maximize the objective function increases exponentially in the number of transceivers. An exact solution to the unconstrained problem is therefore intractable. To illustrate, consider the case of placing three transceivers on a DTED level 2 grid. For the Killeen, Texas map, the DTED grid dimensions are  $901 \times 901$ . Appealing to the theorem, the time complexity required to place three transceivers on the grid is  $O[901^6] \sim O[10^{18}]$ . For each of the  $10^{18}$  configurations of transceivers, the objective function at [2] must be computed and compared.

To be practical, one must resort to heuristic techniques to obtain an approximate solution. One such technique is simulated annealing [K1]. Simulated annealing is a computer optimization technique which models the annealing process of metallurgy, in which a physically stressed metal is first heated, and then cooled at a certain rate to produce a stronger metal. The strongest such metal obtainable by the annealing process corresponds to the global maximum of the objective function formulated for a specific optimization problem.

## The Sensor Placement Analyzer.

Simulated annealing has recently been successfully adapted to the bounded resource allocation problem, in a system called the Sensor Placement Analyzer (SPA) [P1]. The SPA work was sponsored by the Department of the Army with Small Business Innovative Research (SBIR) funding. The objective of SPA is to maximize an objective function corresponding to global visibility (with respect to optical field of view) among cooperative sensor resources. SPA as originally implemented was adapted to a 1:24,000 scale USGS map, representing a 30km x 30km section of Madison County, Virginia. Because of the prohibitive number of sensor placement locations required to address the unconstrained problem, the SPA implementors elected to identify the highest elevation cell within each of nine hundred grid cells (each one km on a side), and restrict sensor placement to these local maxima. The maxima were extracted by hand-selecting contour data, while utilizing a digitizing tablet georeferenced with a scanned map image of Madison County. For each of the nine hundred local maxima, a field of view bitmap was precomputed, depicting area coverage from the site. To constrain the sensor placement problem, a user is able to interactively create polygonal regions corresponding to both red (adversarial) and blue (friendly) areas of interest, which govern respectively where red or blue sensors may be placed. The user is also able to specify with a pull-down menu which of two optimization regimes to run - a simulated annealing version, or a faster locally constrained scheme. Another constraint on the system allows the user to specify whether radio netting constraints are to be obeyed when placing sensors. Although there are no benchmark ground truth data available to test the quality of the system's performance, it is readily apparent to an observer that excellent solutions are derived by SPA.

There are some limitations to SPA, as originally implemented. One is confinement of sensor placement to local maxima, which implies that field of view from local minima (valleys) is not considered. A new version of SPA, tentatively called the general placement analyzer (GPA), addresses this limitation [B2]. Another limitation suffered by any algorithm utilizing simulated annealing is the development of an annealing schedule which lends itself to both high performance and an admissible solution (i.e., an optimal solution). The primary complaint about simulated annealing is usually directed at its slowness to converge. GPA plans to overcome this limitation by utilizing as many domain-specific constraints as possible to reduce the search space to a minimal covering set of potential sensor sites. One constraint which offers good leverage specifies that a collection asset such as a transceiver must be located probabilistically within a relatively tight spatial bound about a roadway.

## Exploiting the fringes of invisible areas to facilitate combined field of view.

Transceiver placement algorithms, even when equipped with field of view bitmaps, are frequently in a quandary regarding where to place a second transceiver after one transceiver has been placed and its field of view displayed. The author suggests the following technique - *place the second transceiver upon the edge of an invisible area*. The edges of invisible areas are often ridges or lips of depressions, offering excellent vistas of regions not visible from the first transceiver location. The order of placement is an open problem. If one adopts a greedy technique, which is not guaranteed to produce an optimal solution, one can select the field of view bitmap with greatest visibility ratio as a good site for the first transceiver. To facilitate cooperative line of sight, one then moves another transceiver to the fringe of the largest invisible region. An alternative is to move a relay to the near fringe and a transceiver to the far fringe. This logic is iterated until the supply of transceivers is exhausted, until the summed visibility ratio approaches one, or until it is observed that no improvement is forthcoming. In Figure 4, if one transceiver is placed at site T, then other transceivers and relays may be placed as shown to enhance combined field of view. The fringe exploitation technique does have limitations, which include: an invisible area may be within hostile territory, there may not be easy transport to the fringe area, or the fringe area may reside beyond electronic line of sight of the first location.



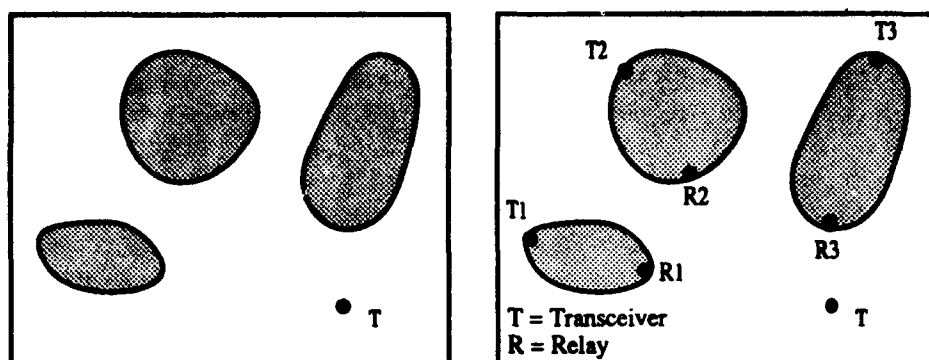


Figure 4. Assigning relays and transceivers to fringes of invisible areas.

### Constraints on IEW asset management.

The constraint set for IEW asset management consists of a variety of databases, some static and some dynamic. Static databases include: a table of the performance characteristics of the transmitters and the receivers under consideration; a set of feasible collection sites based on Digital Terrain Elevation Data (DTED); and a database comprised of the Defense Mapping Agency's thematic vector overlay Tactical Terrain Data (TTD) [M1]. The latter database is projected to contain a set of topographic contours, with contour intervals specified by the map resolution (e.g., a 1:50,000 scale map uses a 10 meter contour interval). Also important are the optical and electronic line of sight databases, which encode fields of view from arbitrary locations. These databases are normally computed in a preprocessing step, which consumes quintic processing time and quartic memory to produce  $n^2$  exact fields of view. Dynamic databases include (but are not limited to): a user-provided representation of the forward edge of battlefield activity (FEBA) and forward line of troops (FLOT); phase lines corresponding to the division area of interest (DAI) and named areas of interest (NAI); no-go or slow-go areas represented by a modified combined obstacles overlay (MCOO).

### Transceiver placement confined to local elevation extrema.

Because of the large number of potential transceiver locations on a map  $n$  grid cells on a side, many approaches to generation of field of view, including the original version of the Sensor Placement Analyzer described above, espouse selecting potential sensor sites based on local elevation maxima. Therefore, the local maxima of the DTED database are suitable candidate locations for transceiver placement. Relatively high elevations generally provide high visibility vistas, making this a powerful technique. However, low spots of the terrain are not accommodated, so it is not possible to develop a field of view from a river valley. It has been observed that certain river valleys, such as the Chosin reservoir valley in Korea, offer relatively wide fields of view [R1]. It is also true that in mountainous terrain, there are many more roads in the valleys and passes than on the peaks themselves, and transceivers are frequently deployed within a relatively tight spatial neighborhood about a road. As a final note, there are many sites on a map which are not local maxima or minima, yet still provide good directional vistas. Shelves or ledges on mountain sides frequently offer excellent one hundred eighty degree vistas of a terrain, but are overlooked during the automated site selection process. When looking for adversarial transmitters, one needs good field of view in the direction of the FEBA. One research effort has devised a color-coding scheme to signify the relative visibility of terrain to a user, based on the user's requirements [R1].

## **Filtering out sites which are too near map features.**

To minimize interference and to avoid detection, Army doctrine specifies that communications equipment be stood off specified distances from map features such as roads, streams, and power lines. If a candidate transceiver location is too near a map feature, then it may be removed from the candidate list of transceiver locations, at the discretion of the analyst. An alternative is to perturb the location slightly, until a new site is generated to meet the requirement. This latter action, however, often relocates the site to a lower elevation, where it may not be appropriate to place a transceiver because of poor field of view.

Previous research at the Intelligence Electronic Warfare Directorate resulted in a set of high-performance algorithms designed to compute the distance from a query point to a set of map features [C1]. These or similar algorithms may be utilized in a preprocessing step to remove those candidate sites deemed too near map features. Alternatively, if one desires to keep the sites, the failed proximity constraints notwithstanding, then each site may be tagged to indicate noncompliance with doctrinal proximity specifications.

## **The division area of interest and forward edge of battlefield activity constraints.**

During planning operations, an Army division is assigned a division area of interest (DAI). Doctrinally, a DAI is a rectangular region of about 20km X 30m, although the shape and dimensions are dependent upon map context. For the asset management problem, the DAI frontage boundary may be used to filter potential transceiver placement locations. The DAI frontage boundary may be viewed as a line segment which segregates potential red transceiver locations from blue.

During conflict, there is sometimes an implicit hypothetical boundary visualized or drawn by an analyst between the frontages of friendly and opposing forces. This boundary, called the forward edge of battlefield activity (FEBA), imposes yet another constraint on friendly transceiver placement. The FEBA may be perceived as a context-sensitive perturbation of the front edge of the DAI. Generally, it is ill-advised to place a transceiver on the side of the FEBA occupied by an opponent, so that only those candidate sites in friendly territory need be considered. Once again, an analyst may override the DAI or FEBA restrictions in the same way he may override a proximity constraint, but in most cases he will find it prudent to allow the constraint to stand. If he does relax the constraint and allows the sites to persist in the database, then the asset management system may tag the sites as having failed the DAI or FEBA constraints.

## **Computing the hearability graph.**

Once the proximity, division area of interest, and forward edge of battlefield activity constraints (given they are available) have been exploited, the list of candidate transceiver placement sites may be subjected to constraints imposed by radio frequency (RF) power propagation modeling. RF line of sight combines power loss with optical line of sight. Using RF propagation models, one may determine if a candidate site is within a specified power contour radially produced from another location. A power contour is derived by performing an RF line of sight computation from one location to another, and the contour is labeled with the signal to noise ratio, expressed in decibels (dB), of a hypothetical communication. Running the RF propagation model over all sites in the candidate list produces a graph. The nodes in the graph are candidate site selections, and an edge represents certitude that a node can hear another node at a specified power level. Observe that a node may represent an entire region of RF reception, and is not necessarily just one point on the map. If two nodes are connected by an edge, then they can communicate with each other at a specific power level, which means that one can hear the communications of the other. Conversely, the lack of an edge between nodes indicates that

there is no network capability between them. For descriptive purposes, we call the resultant connected graph a transceiver hearability graph. In one such graph at Figure 5, potential site locations are linked to other hearable locations, at various power levels.

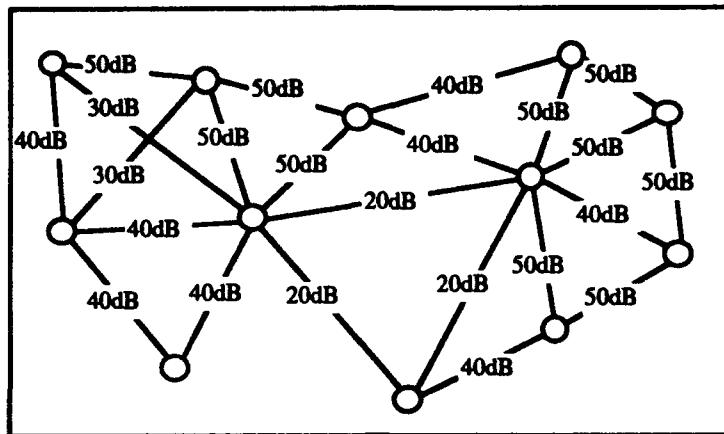


Figure 5. A transceiver hearability graph.

### Uninformed surveillance.

If there is no information available concerning the locations of red transceivers, then blue forces can adopt one of a variety of strategies to become informed. One strategy is to place  $k$  transceivers on the hearability graph at roughly equal distances across the frontage of the DAI, while utilizing transceivers in receiver mode. This is somewhat analogous to a zone defense employed by professional football teams. Another strategy is to use an unconstrained optimization technique such as simulated annealing to find the best optical field of view locations for blue transceivers, place transceivers at those locations, and trust that red transceivers do not deploy in invisible areas. Yet another strategy is to go into surveillance mode with the transceiver capable of acquiring the best field of view, usually an airborne platform, and scan the DAI until activity is detected. Once activity is detected, blue forces may begin to deploy transceivers to populate the hearability graph, motivated by the requirement to monitor red activity.

### Hypothesizing a communications network of opposing forces.

With the DAI, FEBA, hearability graph, and other exploitable constraints in hand, one may proceed to hypothesize the deployment of a communications network managed by opposing forces. If an opponent is responsible for managing  $k$  transceivers, then the  $k$  devices must populate the hearability graph in such a way as to define a network on the hostile side of the FEBA. No single red transceiver is permitted to be outside the communications range of every other red transceiver. Another way of expressing this is that at least one other transceiver must be netted to a given transceiver.

The suspected locations, if they are available, of opposing force transceivers, are plotted on the hearability graph. If a suspected location does not correspond to a node of the hearability graph, then it may be installed into the graph at this time, its RF power propagation contours computed, and its power links connected to hearable red transceivers. In Figure 6, locations of three red transceivers have been hypothesized, indicated by the shaded nodes. The transceivers have been labeled RL, RC, and RR, which represent red's left, center, and right respectively. Note that RL is netted to RC, which is in turn netted to RR.

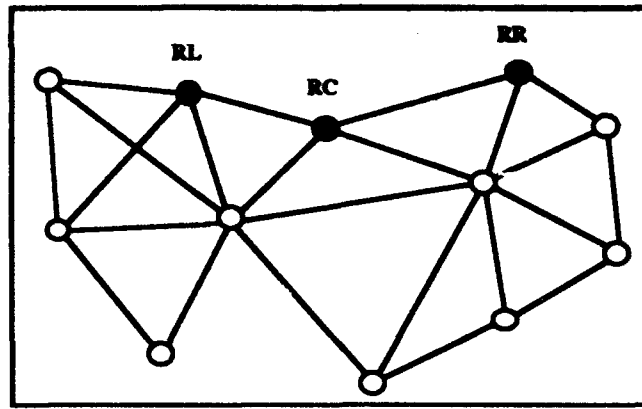


Figure 6. A simulated red communications network deployment.

### Informed surveillance and parity matching.

If and when suspected or hypothetical locations of red transceivers become available, due to information gathered during the intelligence cycle, it becomes worthwhile to attempt to identify optimal collection sites for blue transceiver placement. These sites must obey the RF propagation power constraints imposed by the hearability graph.

*Parity matching* is used to implement a one-on-one assignment of blue to red transceivers, with an accompanying abandonment of the zone defense occasioned by uninformed surveillance. Parity matching is a technique to counter every located adversarial transceiver with a friendly one. The surveillance regime transfers from uninformed to informed. The underlying assumption is that for every red transceiver deployed, there is a corresponding blue transceiver available to counter it. In Figure 7, it has been determined that to cover red's center transceiver effectively, there are two possible locations, BC1 and BC2, for blue's center transceiver. It is computationally more efficient to begin the allocation process with blue's center rather than the extremes of the network.

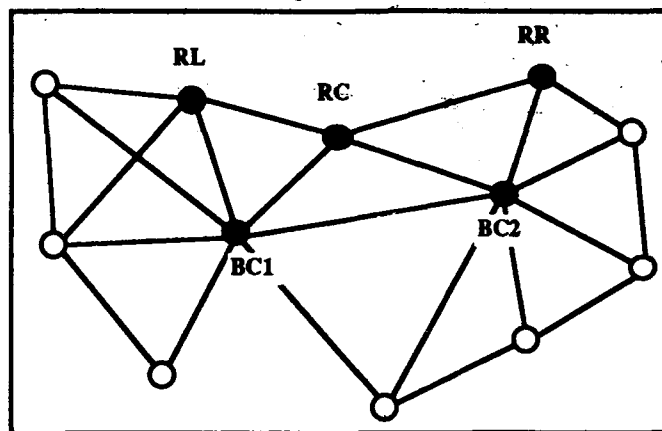
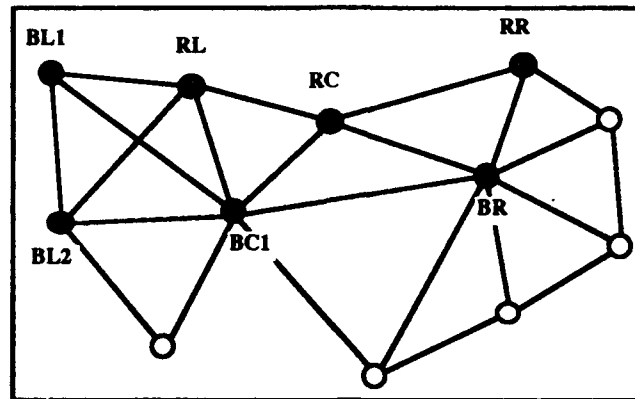


Figure 7. Parity matching: lining up blue's center with red's center.

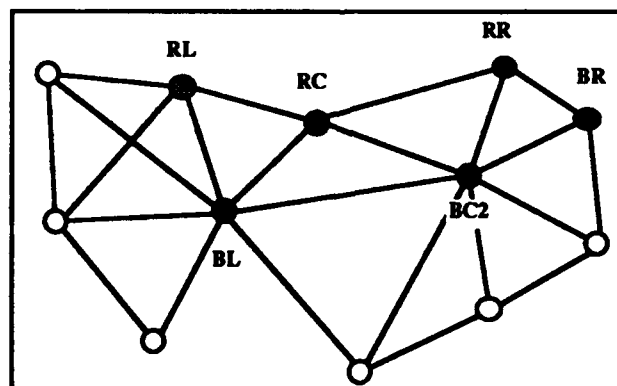
It is now feasible to develop the topology (graph-theoretic deployment) of blue's network. If location BC1 is selected to be the location of blue's center, then there are two possible locations

(connected to red's left) for blue's left, but only one for blue's right. Thus there are two feasible transceiver network deployments using site BC1 as blue's center (Fig. 8). For blue's left, one might select one of the two available locations BL1 or BL2 based on stronger signal-to-noise ratio, better logistics support, or a variety of other operational constraints. Note that with this topology, blue's left provides redundant coverage, since blue's center alone is capable of monitoring both red's left and center.



**Figure 8. Possible locations for blue's flanks, given BC1 center selection.**

If site BC2 is selected as blue's center location, then there is a unique allocation strategy, which puts blue's left at BL and blue's right at BR (Fig. 9). In this instance, blue's right may be detached from the network since blue's center transceiver covers both red's center and right. When this allocation is added to the two found previously, it is seen that there are three feasible transceiver allocation strategies for this small example. In practice, for large scale problems, there are likely to be many allocation solutions which satisfy a commander's transceiver allocation requirement.



**Figure 9. Possible locations for blue's flanks, given BC2 center selection.**

### **Totally constrained transceiver allocation and the stable marriage problem.**

In this section, the algorithmic complexity of the totally constrained transceiver allocation problem is discussed. The allocation problem is totally constrained when both the hearability graph is available and the locations of adversarial transceivers are known. It will be seen that these constraints make the allocation problem much simpler than unconstrained optimization. However, it should be emphasized that locational constraints are available only infrequently, during regimes of

informed surveillance.

**Theorem.** The totally constrained transceiver allocation problem is equivalent to the stable marriage problem.

**Proof:** In the stable marriage problem [K2], it is required to match  $k$  bachelors with  $k$  maidens in such a way that the resulting marriages are stable. By "stable", it is meant that there do not exist "more suitable" marriages between spouses which might cause the given set of marriages to dissolve. To illustrate, two couples who prefer their own spouses to those of another couple have stable marriages. Turning to the transceiver allocation problem, it may be construed that blue receivers correspond to bachelors, red transmitters to maidens, and signal-to-noise-ratio between blue and red to the marriage "bond". It is possible to stabilize an unstable transceiver allocation by requesting two blue transceivers to swap surveillances. The allocation of Figure 10 (matching is bold) is unstable, since it is possible to find another matching with greater signal-to-noise ratio (the 50 and 40 dB links). To stabilize the allocation, one simply needs to detach the crossing 20 dB links and enable the 50 and 40 dB links.

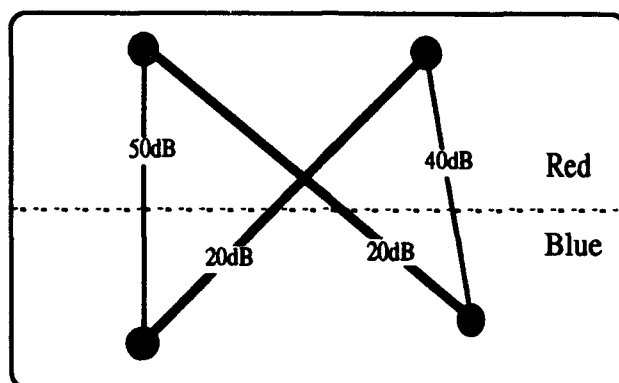


Figure 10. An "unstable" transceiver allocation.

**Theorem.** The totally constrained bounded resource problem is solvable in linear time.

**Proof:** This follows immediately, since the stable marriage problem is solvable in linear time [S2].

### Unconstrained vs. constrained optimization: computing the savings.

It was shown above that in an unconstrained application the time complexity to place  $k$  transceivers on a gridded map of dimension  $n$  is  $O[n^{2k}]$ . It was also shown that for a totally constrained application, the time complexity is  $O[k]$ . The savings in performance is the difference between these two functions. As the number of transceivers grows, the total savings is the integral of the difference of the functions. The savings itself is exponential in complexity, as represented at equation [7] below.

$$t = a_1 n^{2k} + b_1 \quad [4]$$

$$t = a_2 k + b_2 \quad [5]$$

$$A = \int (a_1 n^{2k} + b_1 - a_2 k - b_2) dk \quad [6]$$

$$= \frac{a_1}{\ln n} n^{2k} - \frac{a_2}{2} k^2 + b_3 k + c \quad [7]$$

## Specificity of constraints and algorithmic complexity.

It has been demonstrated that in an unconstrained transceiver placement application, the time complexity to solve the allocation problem is exponential in the number of transceivers, while in a strongly constrained regime, the complexity is linear in the number of transceivers. Although open problems remain for sets of weaker constraints which lie somewhere between these extrema (see tabulation below), it is surmised that the time complexity for the weakly-constrained problem is bounded between exponential and linear. For example, when only the hearability graph is available, but the locations of adversarial transmitters are unknown, the domain is similar to the quadratic matching problem, which has time complexity  $O[n^2]$ .

### UNCONSTRAINED OPTIMIZATION $O[n^{2k}]$

\*\*\*\*\*

1. Transceiver sites permitted anywhere on map  
:
2. Transceiver sites confined to local maxima  
*Filter sites based on local maxima*  
:
3. Proximity to interfering map features is known  
*Filter sites based on proximity/inclusion*  
:
4. Division area of interest (DAI) known  
*Filter sites based on DAI frontage*  
:
5. Forward edge battlefield activity (FEBA) known  
*Filter sites based on FEBA frontage*  
:
6. Optical line of sight bitmaps available for some locations
7. Electronic line of sight bitmaps available for some locations  
*Compute partial hearability graph*  
:
8. Optical line of sight bitmaps available for all locations
9. Electronic line of sight bitmaps available for all locations  
*Compute complete hearability graph*  
:
10. Location of some red transceivers suspected  
*Find possible blue assignments*  
:
11. Network topology of two or more red transceivers suspected  
*Partial one-on-one blue/red matchings*  
:
12. Location of all red transceivers suspected
13. Network topology of all red transceivers suspected  
*One-on-one parity matching of blue on red*  
*Unstable matching (weak S/N ratios permitted)*  
*Stable matching (optimal S/N ratio matching plotted)*

\*\*\*\*\*

### SEVERELY CONSTRAINED OPTIMIZATION $O[k]$

## Massive parallelism: the proposed tByte-LoSight machine.

When developing an ensemble of field of view bitmaps corresponding to specific sites within a gridded database, it is possible to gain algorithmic performance by utilizing concurrency wherever possible. The proposed tByte-LoSight machine is a massively parallel architecture concept designed to perform field of view bitmap constructions and cooperative transceiver placement. The machine as envisioned is dimensioned as a  $1000 \times 1000$  array of processors, with each processor having access to one megabyte of local RAM. The million processors have a combined access to a million bytes of local RAM - hence the tByte notation.

*Field of view calculations.* A processor's location in the tByte-LoSight array is analogous to the corresponding location in Digital Terrain Elevation Data, Digital Elevation Matrix, or similar gridded array. Each processor's RAM is preloaded with the values from the gridded database. Then, in parallel, the field of view bitmap  $\lambda_{i,j}$  for each location  $(i,j)$  is produced for each element of the array, and is used to overwrite the grid data in local RAM. With this architecture, tByte-LoSight is capable of storing up to a million bitmaps, each one megabyte in size. Using an approximate field of view algorithm, to produce an ensemble of  $n^2$  bitmaps, the performance of the architecture is  $O[n^2]$ . An alternative, more finely-grained architecture replaces each processor with an  $n \times n$  array of processors, which speeds up the bitmap generation process to  $O[n]$ . The performance improvement is due to each low level processor computing a single LOS radius rather than  $n^2$  radii.

*Transceiver placement.* When during asset management processing a transceiver is tentatively placed at cell  $(i,j)$ , the tByte-LoSight machine fetches bitmap  $\lambda_{i,j}$  and inverts it to produce the complement of  $\lambda_{i,j}$ , denoted  $\sim\lambda_{i,j}$ . Concurrently, tByte-LoSight then computes the set difference between each of its elements and  $\lambda_{i,j}$ , followed by the set difference between  $\sim\lambda_{i,j}$  and the result, while progressively updating local RAM. Finally, a logical operation is performed to find the processor whose RAM most closely resembles the null bitmap, for this location is deemed to be the best site for a second transceiver. From the second site's location, this logic may be iterated until the summed visibility ratio approaches one, until the supply of transceivers is exhausted, or until it is judged that no further improvement is possible. The tByte-LoSight architecture is currently just a concept. The cost and reliability of such a massive architecture are issues, as are data transfer rates, power requirements, and communications polling costs.

## Conclusions.

It has been shown that the Intelligence Electronic Warfare transceiver allocation problem exhibits a wide spectrum of algorithmic complexity, depending upon the availability of constraints to bound the problem. At one extreme, when no constraints are available, the time complexity required for an optimal field of view solution is exponential in the number of transceivers allocated. On the other hand, for the tightly constrained problem, when the radio frequency propagation graph is known, along with suspected locations of the transmitters of interest, the time complexity is linear in the number of allocated transceivers, since the problem reduces to Knuth's stable marriage theorem. It is speculated, but has not yet been proven, that when only weaker constraints are available, the complexity falls somewhere in between. Another result to emerge from this research is a new technique to place transceivers on the fringes of areas invisible to an already placed transceiver. Finally, a massively parallel architecture has been conceptualized to support both field of view and transceiver placement, although infeasible power requirements preclude implementation in the short term.



## Acknowledgements.

The Jarrett-Riding study is an excellent overview of the state of the art in field of view algorithms. The paper by Pittard et. al., together with an accompanying computer demonstration, is the most insightful work that the author has encountered dealing with the transceiver placement problem. Figures 1 and 3 were generated with the Plot3D and ContourGraphics commands of Wolfram Research's Mathematica, version 2.0, using the function  $\sin[x] \sin[y]$ . The research benefited from discussions with J. Allen, G. Andersen, R. Antony, J. Benton, D. Brown, R. Butler, D. Chubb, K. Clark, J. Cooley, B. Cummings, J. Dennis, J. Jarrett, K. Langdon, J. Mitchell, L. Pittard, C. Ray, D. Sappington, and M. Sweedler.

## References.

[B1] Broome, B., *SEEFAR: An Improved Model for Producing Line-of-Sight Maps*, Army Material Systems Analysis Activity (AMSAA) Technical Report no. 225, Aberdeen Proving Grounds, September 1980.

[B2] Brown, D., private communication.

[C1] Cronin, T., *High Level Design Specification for Proximity / Inclusion Algorithms Adapted to the Defense Mapping Agency's Tactical Terrain Data*, US Army CECOM RDEC Intelligence Electronic Warfare Directorate Basic Research Transition document, Vint Hill Farms Station, Warrenton VA, December 1992.

[J1] Jarrett, J., and T. Riding, *Line-of-Sight-Analysis Study FY92*, US Army Topographic Engineering Center Technical Report, Ft Belvoir VA, 1992.

[K1] Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi, *Optimization by Simulated Annealing*, Science, 220(4598):671-680, 13 May 1983.

[K2] Knuth, D., *Marriages Stable*, Les Presses de l'Universite de Montreal, Montreal, 1976.

[M1] Messmore, J. and L. Fatale, *Phase I Tactical Terrain Data (TTD) Prototype Evaluation*, US Army Engineer Topographic Laboratories Technical Report ETL-SR-5C, Ft. Belvoir VA, December 1989.

[P1] Pittard, L., D. Brown, and D. Sappington, *SPA: Sensor Placement Analyzer, an Approach to the Sensor Placement Problem*, Proceedings of the 10th Annual JDL Conference on Command and Control Decision Aids, National Defense University, Washington DC, July 1993.

[R1] Ray, C., *A New Way to See Terrain*, unpublished manuscript, US Army Military Academy, April 1993.

[R2] Ray, C., private communication.

[S1] Sciandra, R., *TIREM/SEM Programmer's Reference Manual*, Department of Defense Electromagnetic Compatibility Analysis Center, Technical Report ECAC-CR-90-039, Annapolis MD, July 1990.

[S2] Sedgewick, R., *Algorithms in C*, Addison-Wesley, Reading MA, 1990.

[V1] *Vector Smart Map Level 2 Databases*, Military Specification MIL-V-89032, Defense Mapping Agency, January 28 1993.

# **CEMPLOT: A VISUALIZATION TOOL FOR CEM VII**

**Kiran Shivaswamy, Patrick Burns and David Alciatore**

**Department of Mechanical Engineering**

**Colorado State University**

**Fort Collins, CO 80523**

**kira@carbon.LANCE.ColoState.EDU,**

**pburns@westnet.net,**

**alciator@longs.LANCE.ColoState.EDU**

**ARO Contract #DAAL03-92-G-0176**

## **ABSTRACT**

The Concepts Evaluation Model VII (CEM VII) is the latest version in the progression of US Army theater force model development. It has been developed under the auspices of the US Army Concepts Analysis Agency (CAA). The model is used by the Army to assess and optimize combat force capability. Typically, the model involves battle simulations of many months duration, encompassing multiple armies and diverse terrain.

CEMPLOT is a visualization tool used to portray the results of battle simulations performed in CEM VII. It fulfills the need for a user-friendly, interactive graphics tool with different display options. The interface is developed to run in the X Windows and OpenWindows environments. The program is written in C and it utilizes Xlib and XGL libraries. The user selects various frames of information from a menu, developed using the XView toolkit.

Some of the options offered to the user are: 1) display of the battle terrain, 2) display of the location of the Forward Edge of the Battle Area (FEBA) of the two conflicting forces at different time steps, 3) display of resources such as number of artillery pieces and close air support units for the assembled army units, 4) display of the names of army units in combat and 5) information regarding the personnel and ammunition losses incurred by the two opposing forces at different time intervals.

# **1. INTRODUCTION**

## **1.1 Historical Perspective**

The Concepts Evaluation Model VII (CEM VII) is the discrete event battle simulation model, which exists under the auspices of the US Army Concepts Analysis Agency (CAA), located in Bethesda, Maryland. CEM VII is used by the Army to assess and optimize combat force capability. The kernel of CEM VII capability is the ATCAL (Attrition Model Using Calibrated Parameters) algorithm [Johnsrud, 1980] which is used to perform engagements. CEMPLLOT is the visualization tool developed at the Department of Mechanical Engineering, Colorado State University, to display and interpret information produced by CEM VII. Typically, CEM VII involves simulations of many months duration, encompassing multiple armies and diverse terrain. The utility of the simulation lies in the degree to which the information produced by one or more runs of the model can be interpreted. A PC plotting program [Stoll, 1992] exists, which visualizes CEM VII output data. CEMPLLOT, besides duplicating the functionality of the PC program in a Sun X environment, presents the user with additional features to help visualize CEM output.

## **1.2 CEMPLLOT Program Overview**

CEMPLOT runs in the X Windows and OpenWindows environments and uses the XGL graphics library extensively (XGL 2.0). It uses a graphics toolkit XView to provide user-interface objects such as menu buttons. The program has the capability to display CEM output data in color and accepts information and commands from the user, interactively. The user-interface consists of a base frame or window which contains the graphics canvas (e.g., where the terrain and FEBA are plotted) and the input panel (control buttons). The button items include panel buttons which perform a function when clicked on, and menu buttons which allow the user to select a particular option from a menu.

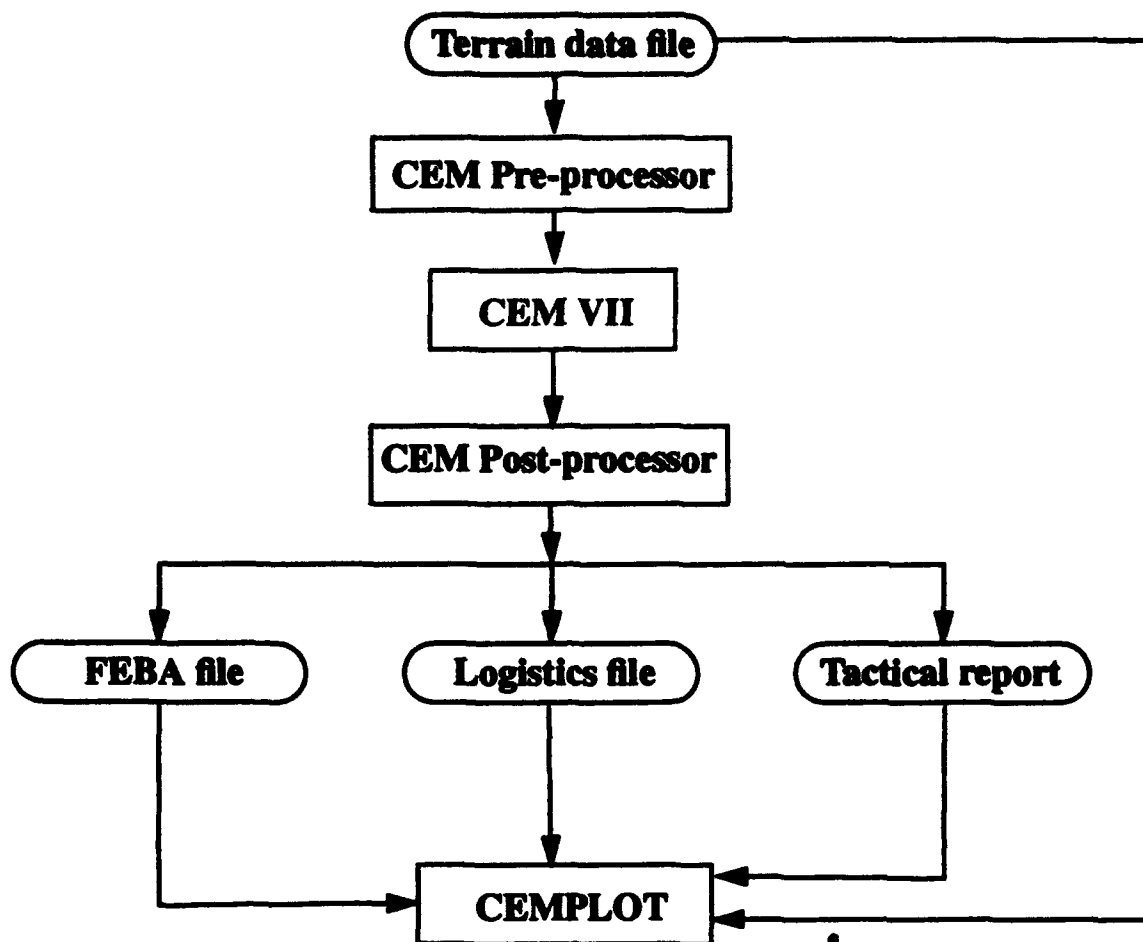
The position of CEMPLLOT in the CEM hierarchy is displayed in the flow-chart of Figure

1, showing the data files which form the input to the visualization tool. The program exists as a collection of C routines which implement the various options of CEMPLLOT. The program has the capability to:

- (i) Display the battle terrain.
- (ii) Display the Forward Edge of Battle Area (FEBA), overlaid on the terrain, for every theater cycle.
- (iii) Display tactical report data, such as number of personnel, artillery, close air support units, etc., for each unit assembled along the FEBA over different theater cycles.
- (iv) Plot theater cycle number versus logistics data, such as amount of ammunition available, versus time step.
- (v) Present a roll-call of all units assembled on the opposing sides at a particular theater cycle (time step).
- (vi) Perform geometrical transformations (scaling and rotation) on the FEBA/terrain plot.
- (vii) Animate the sequence of FEBA movement.

## **2. CEMPLLOT: INPUT FILES**

CEMPLOT uses many input files, one of which is mandatory. The mandatory input file to the visualization program is a battle terrain data file. This file is one of the input files for the pre-processor which is input to CEM VII (refer Figure 1). The other input files include FEBA, tactical report, and logistics data files, which are output from the CEM VII post-processor [Allison et al., 1985]. The following sections discuss the data contained in these files in greater detail.



**Figure 1 Position of CEMPLOT in the CEM hierarchy**

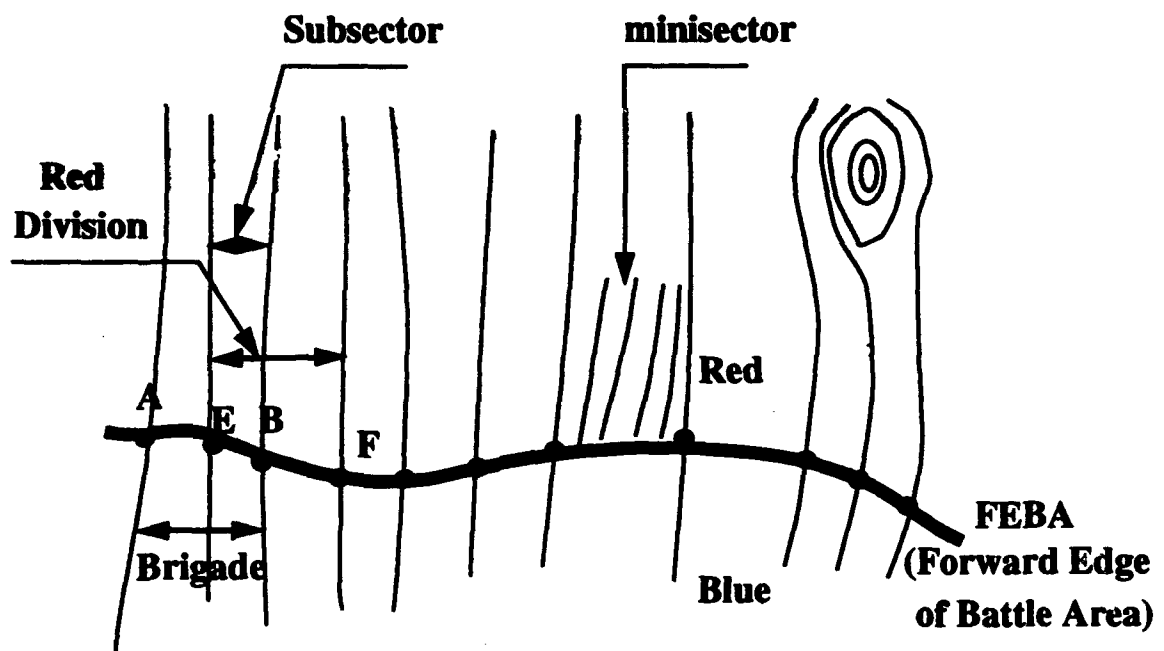
## **2.1 Battle Terrain Data**

The CEM uses a simplified representation of the battlefield [Johnson, 1985]. The distances in the direction of troop movement (vertical) are in kilometers with a minimum distance of one-tenth of a kilometer. Distances along the front (horizontal) are measured in minisectors (see Figure 2). The battle terrain is diverse in nature and is categorized as types A, B, C and D. The four terrain types denote the general nature of the terrain and have some bearing on the mobility of ground combat units.

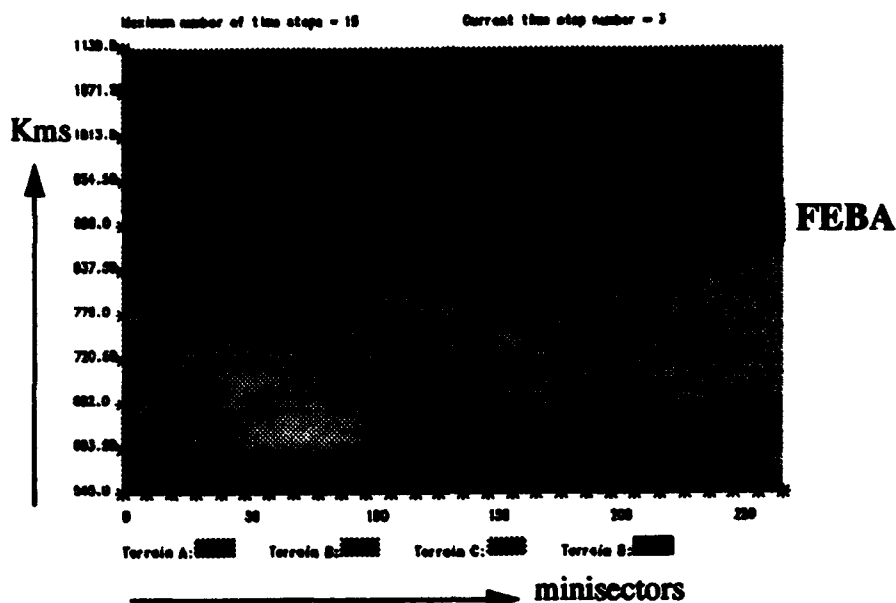
- (i) **Type A:** This terrain is mainly flat and is excellent tank country.

- (ii) Type B: Rolling lands and is marginally suited for tanks and wheeled vehicles.
- (iii) Type C: Mountainous territory. On this terrain, tanks and wheeled vehicles must remain on the roads because of steep slopes.
- (iv) Type D: It represents some major obstacle that would normally require extra or special effort for the forces to negotiate or pass through. It may be a river, lake, marsh, canyon or some man-made barrier, such as a minefield.

The terrain data file contains information regarding number of engagements, number of minisectors and the terrain extents for sets of minisectors. The depth of a barrier of D terrain is set to the minimum distance of movement of one-tenth of a kilometer. For example, a particular terrain band, say, 10 minisectors wide, may be comprised of 3.7 km of Type B, followed by 9 km of Type A, followed by 0.1 km of Type D, etc.



**Figure 2 Unit Deployments (Schematic)**



**Figure 3 Terrain/FEBA Map From CEMPLOT**

## **2.2 Unit Deployment Data**

The deployed forces are resolved to brigade for the Blue side and division for Red. Figures 2 and 3 illustrate the deployment of a Blue force opposing a Red force. A Blue brigade front (such as the distance AB in Figure 2) or a Red division front (e.g. EF) is defined as a sector. The forward line of assembled troops (Blue and Red) is termed the FEBA (Forward Edge of Battle Area). The engagements between the two forces results in the forward and backward movement of the FEBA, with time. The data regarding the position of each minisector along the FEBA, for every theater cycle is contained in a data file, which is output from the CEM VII post-processor. CEMPLOT reads and displays information from the FEBA data file.

## **2.3 Tactical Report Data**

The tactical report contains information, for every  $n^{\text{th}}$  division cycle, regarding the loca-

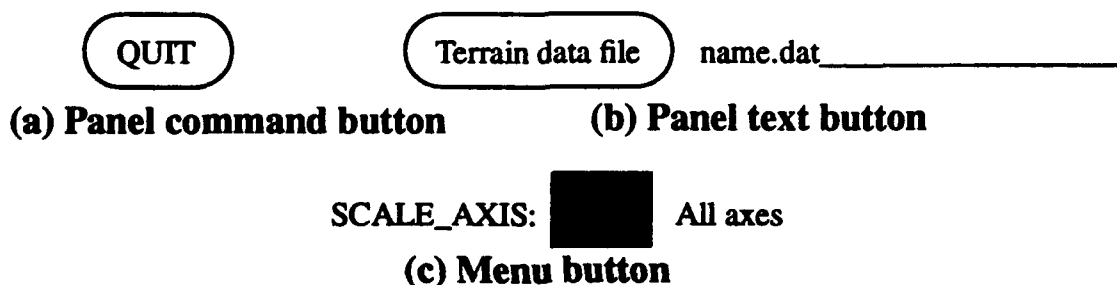
tion, mission, status, authorized troop strength, supplies and number of major weapons assigned to the units in combat. The units, for which the information is provided, include Blue brigades and Red divisions. The unit tactical report generated by the CEM post-processor is passed through a file conversion program to maintain file format compatibility with the existing PC CEM plotting package. The converted file is then passed as input to CEMPLLOT.

## 2.4 Logistics Data

The logistics report presents data concerning the consumption and replacement of resources by the combat units. It includes the total amount of resource lost, both temporarily and permanently, due to both combat and noncombat causes since the start of engagements. The resources, include the number of participating personnel and the amount of ammunition expended. CEMPLLOT reads the logistics data and displays it in the form of Cartesian plots.

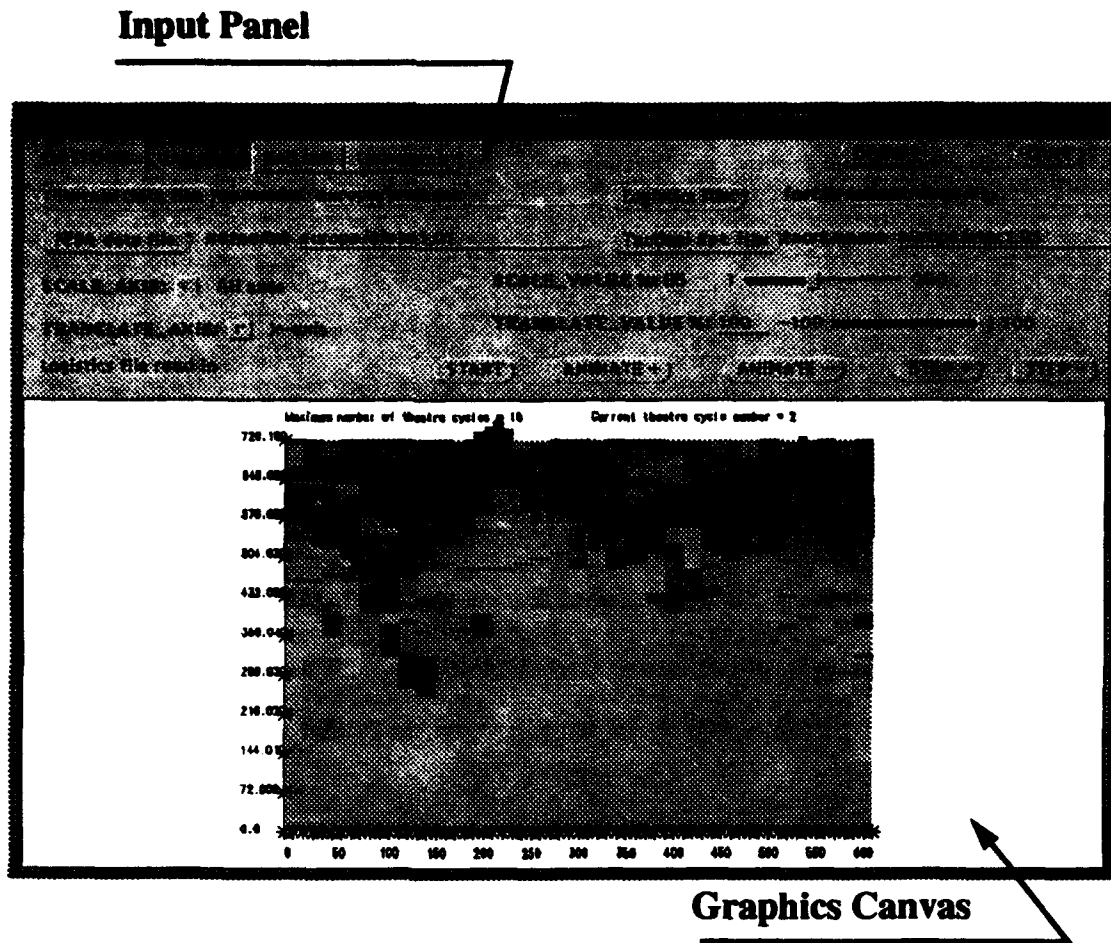
## 3. CEMPLLOT: PROGRAM OPERATION

The program, CEMPLLOT, is run in the UNIX operating system, either in the X Windows or OpenWindows environment. The program, to start with, brings up an application window or base frame as depicted in Figure 5. The base frame has an input panel which consists of command, text and menu buttons, examples of which are shown below (Figure 4).



**Figure 4 Examples of Panel Buttons**





**Figure 5 CEMPLLOT base frame**

The user types the names of the CEMPLLOT input data files in the text space provided to the right of the corresponding panel text buttons. This action is followed by clicking the particular text button using the left mouse button [Shivaswamy and Burns, 1993]. The terrain and FEBA maps are brought up in the graphics area or the canvas of the base frame.

The four different terrain types are represented in CEMPLLOT by different colors. The terrain is displayed laterally as bands, each of which is a pre-specified number of minisectors wide. The forward line of the assembled troops or the FEBA is displayed as a combination of a blue and a red band, each of fixed height (refer to Figures 3 and 5). The individual units or sectors on each side fall on minisector boundaries and are delimited in CEMPLLOT using vertical yellow lines along the FEBA.

## 4. CEMPLOT: USER OPTIONS

These commands are activated by clicking the left mouse button over the required option button. The options available under CEMPLOT are discussed below.

### 4.1 Print Unit Data

"Unit Data" is the default option under the CEMPLOT main menu. This option prints the data regarding number of personnel, supplies and major weapons for the "picked" combat unit. The option brings up the terrain map with the FEBA overlaid on it. The user "picks" a unit by driving the cursor onto that particular unit (between two adjacent yellow lines) and clicking on it using the left mouse button. The unit thus picked is highlighted with an asterisk. The tactical report data pertaining to that unit is displayed in an output window (see Figure 6) which may be moved around or iconified just as any other window. This option retains its capability even after the terrain and FEBA plot is scaled and translated.

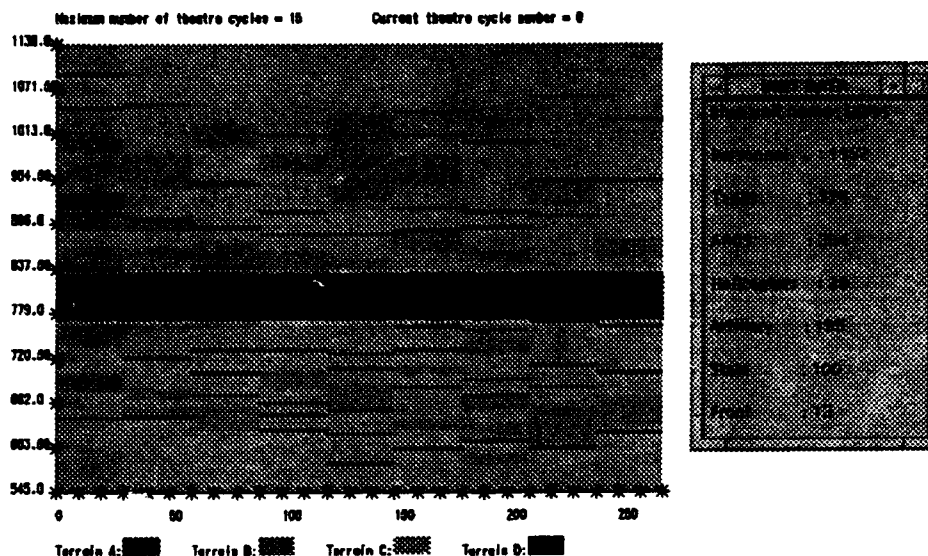
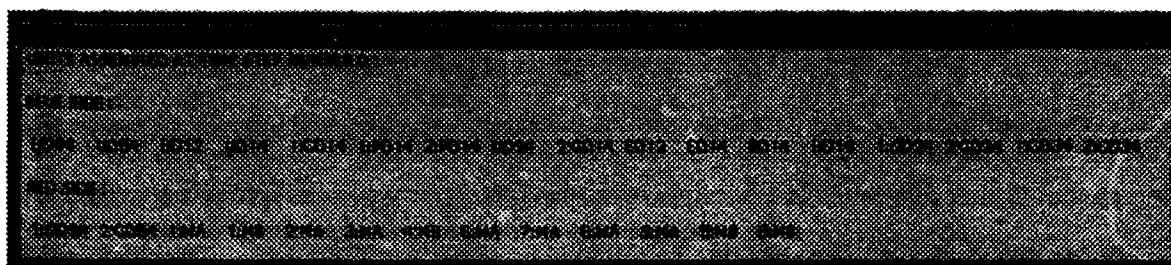


Figure 6 CEMPLOT: Unit Data Display

## 4.2 Print Roll Call

Under this option, the names of all the units present at the end of the current theater cycle are displayed. The data are output, in a manner similar to the previous option, in a new window, as shown in Figure 7. The data displayed include the theater cycle number (time step), the names of all Blue brigades and Red divisions in combat. The output window can be moved around or iconified as in the previous option. The terrain map with the FEBA information overlaid on it will remain displayed in the main graphics window while running this option.



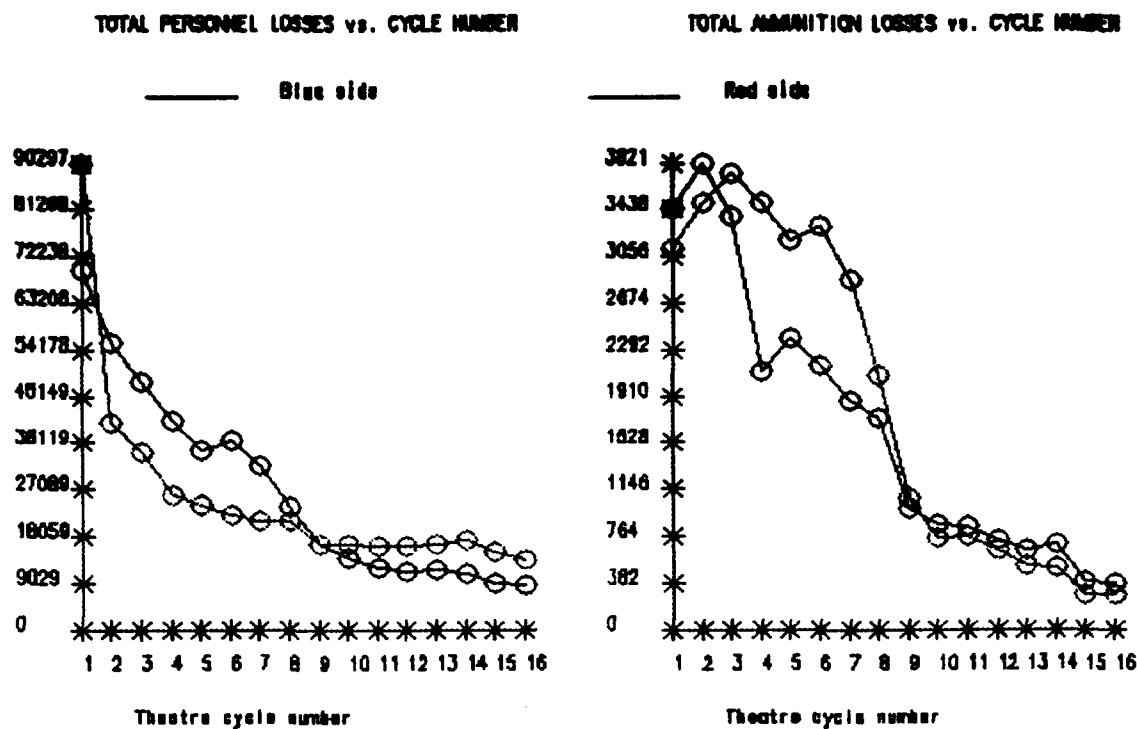
**Figure 7 CEMPLLOT: "Roll Call" Output Window**

## 4.3 Show Logistics vs. Theater Cycle Graph

The total unit losses of personnel and ammunition, at each theater cycle, are plotted versus the theater cycle number in Cartesian form. This includes data for each Blue brigade and Red division. The screen output of this option of CEMPLLOT is shown in Figure 8.

## 4.4 Scaling and Translation

These options allow the user to scale or translate the plot on the screen in X, Y, Z, or uniformly in all the three coordinate directions. The user selects the axis of transformation from a menu and the scaling or translation value from a slider bar. The scaling and translation values are a percentage of the screen plot and are cumulative in nature. Thus, scaling first by 200% and then



**Figure 8 CEMPLLOT: logistics vs. cycle number graph**

by 50% results in a cumulative scaling of 100% (unchanged). These options are particularly useful when many combat units are assembled across a few minisectors, in which case the plot could be scaled to yield more detailed information.

#### 4.5 'Start', 'Step +' and 'Step -' Options

The battle engagements between the opposing forces take place over a specified number of time steps or theater cycles. The above options allow the user to view data present in the FEBA and tactical reports for any theater cycle. The 'Start' button sets the theater cycle number to zero, enabling the user to view information pertaining to start of battle engagements. The options 'Step +' and 'Step -' increment and decrement the FEBA plot by one theater cycle number, correspondingly.

#### **4.6 Animation: 'Animate +' and 'Animate -'**

These options animate the sequence of FEBA movement for theater cycles zero to maximum. 'Animate +' and 'Animate -' move the FEBA from zero to maximum, and from maximum to zero, respectively. These options are helpful as they provide the user with an overview of the progression of the entire battle.

### **5. CEMPLOT: Examples of Use**

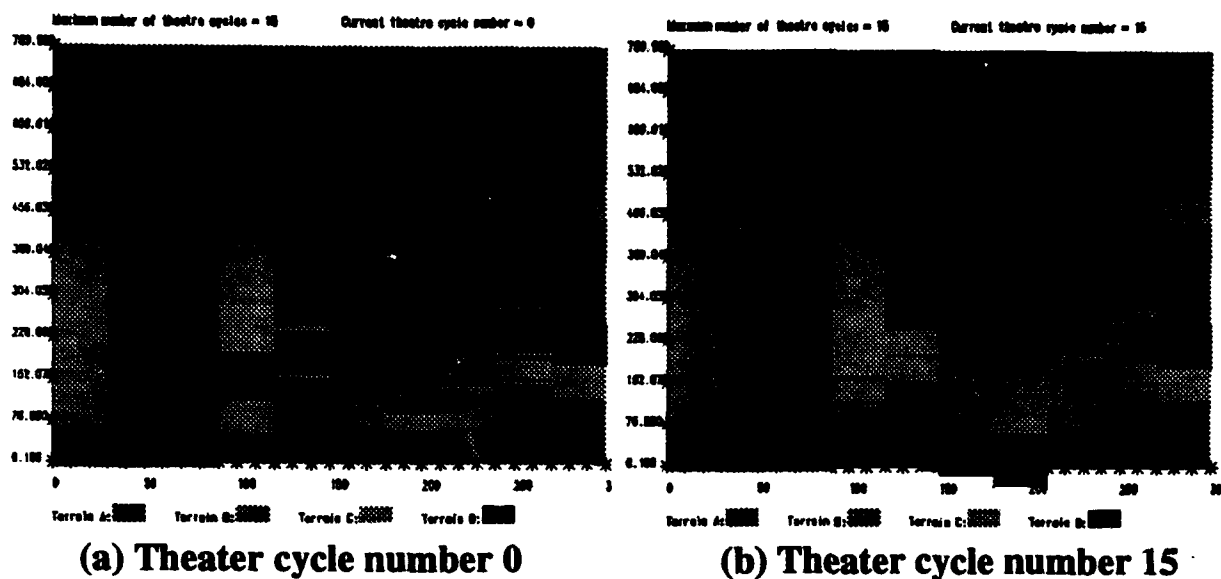
#### **5.1 Interpreting Terrain Features and FEBA Movement**

The terrain and FEBA map displayed by CEMPLOT can be studied for the distribution and extents of each terrain type and their influence on the movement of forces. This can be done very easily by stepping through the different theater cycles and studying the displacement of the FEBA. This study can be repeated for different terrain data sets and their corresponding FEBA location files. Figures 9, 10, and 11, display terrain and FEBA data sets for the Campaign I, Europe, and Ardennes data sets, respectively. Each figure displays the position of the FEBA at the beginning and end of the simulation (in other words, at the end of theater cycle number zero and maximum, respectively).

#### **5.2 Parametric Studies**

CEMPLOT can be used to visualize any CEM VII simulation model across force variations. It provides a user-friendly and robust tool for the visualization of data generated by different runs of CEM VII. CEMPLOT can display several measures whereby one force variation can be compared to another. These include the displacement of the FEBA, the total resources expended by each side, and the time taken to arrive at a particular state or condition of the battle. The user can achieve this very easily by running multiple copies of CEMPLOT (invocation in the back-

ground mode), providing each with data files output from varying input parameters. The separate CEMPLLOT windows thus generated can be scaled and set adjacent to one another for convenience in comparison.

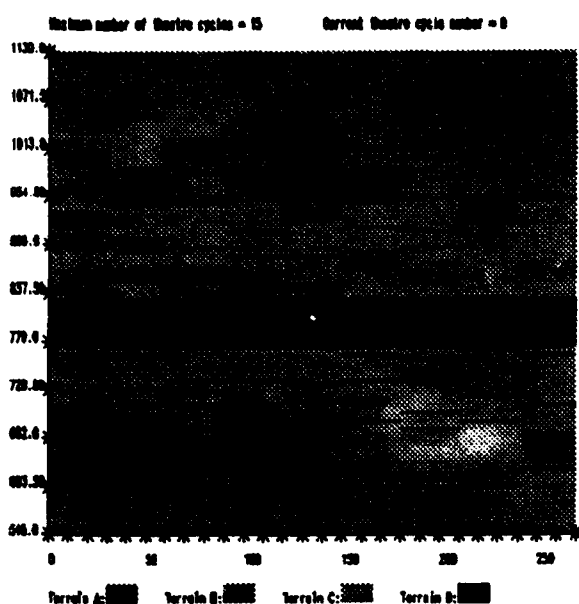


**Figure 9 Campaign I data set**

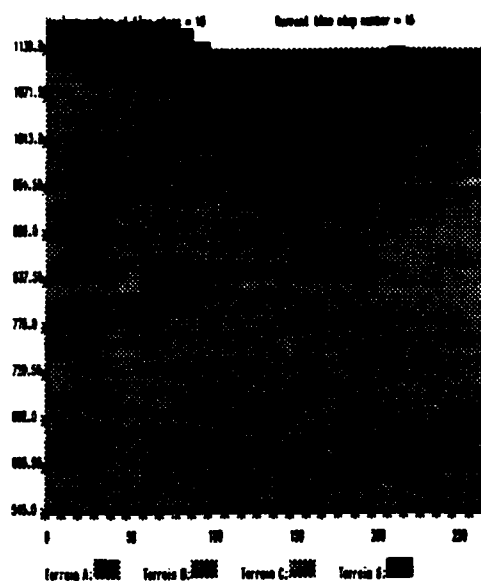
### 5.3 Evaluation of CEM VII Algorithms

CEM VII, including its pre- and post-processors occupy nearly 200 MegaBytes of disk space. The post-processor outputs up to 18 different data files, some of which are in binary format. The data files are very large in size and are daunting to interpret textually. The massive amounts of data generated mandate that a special means of display be developed in order to interpret and display them. CEMPLLOT seeks to fulfill this need, in a user-friendly fashion, in real time, in a UNIX workstation environment.

Efforts at Colorado State University [Brewer and Burns, 1991] have been directed towards the vectorization of the ATCAL algorithm and its insertion into CEM VII. CEMPLLOT has proved

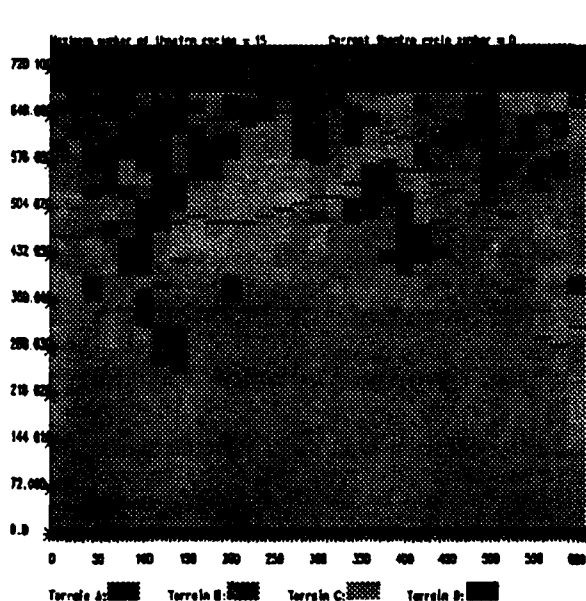


(a) Theater cycle number 0

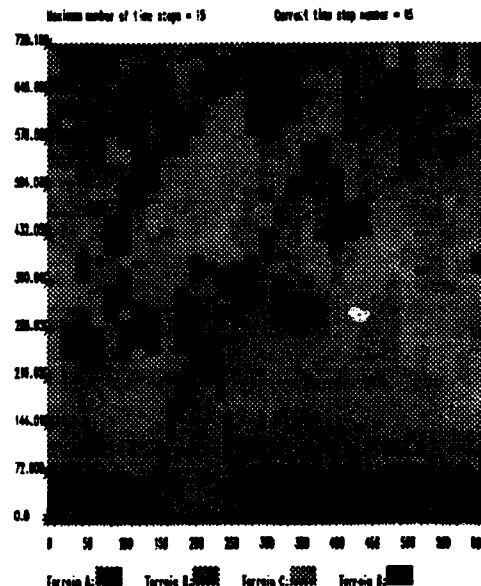


(b) Theater cycle number 15

Figure 10 Europe data set



(a) Theater cycle number 0



(b) Theater cycle number 15

Figure 11 Ardennes data set

to be very useful in interpreting the results of vectorization and in comparing them to those obtained from unvectorized algorithms. The evaluation is done by exploring the convergence of the unvectorized and vectorized codes. The global variation of FEBA with time is adopted as a metric of convergence, which is then studied from the display generated by CEMPLLOT.

## **6. Future Work**

### **6.1 Incorporating Posture Information**

Each combat unit, assigned to Blue brigade and Red division, is associated with a particular posture for the engagement [Allison et al., 1985]. This may be one of the following eight types: (1) Blue attack; Red delay, (2) Blue attack; Red prepared defense, (3) Blue attack; Red hasty defense, (4) Blue attack; Red attack, (5) Red attack; Blue hasty defense, (6) Red attack; Blue prepared defense, (7) Red attack; Blue delay, and (8) Static. Efforts are presently underway to display this information for all the assembled units at the end of each theater cycle.

### **6.2 3-Dimensional Fractal Display of Terrain**

A realistic interpretation of the battle terrain data has been attempted making use of the technique of fractals. This effort involves modeling the terrain data using a fractal generating algorithm in a 3-dimensional format. The generation and display of such a model has been achieved and efforts are underway to incorporate it as an option under CEMPLLOT.

## **ACKNOWLEDGEMENT**

We are very grateful for the assistance of Mr. George Stoll of the US Army Concepts Analysis Agency, who supplied us with the PC visualization package for CEM, and for providing us with information pertaining to our queries.



We are also grateful to Mr. Bill Allison of the CAA, for his assistance with the running of CEM VII and its post-processor. Finally, we are grateful to Mr. Gerry Cooper of the CAA, for his patient and insightful supervision of our efforts.

This work was supported under ARO Grant DAAL03-92-G-0176.

## 7. REFERENCES

Allison, W. T., Devlin, H., and Johnson, R. E., 1985. *Concepts Evaluation Model VI (CEM VI), Volume II - User's Handbook*, US Army Concepts Analysis Agency, CAA-D-85-1, Bethesda, MD, rev. December 1991.

Brewer, M., and Burns, P., 1991. *A Convergence Study of CEM VII*, Internal Publication, department of Mechanical Engineering, Colorado State University, Fort Collins, CO 80523, April, 1993.

Johnson, R. E., 1985. *Concepts Evaluation Model VI (CEM VI), Volume I - Technical Description*, US Army Concepts Analysis Agency, CAA-D-85-1, Bethesda, MD, rev. October 1987.

Johnsrud A. E., 1983. *ATCAL: An Attrition Model Using Calibrated Parameters*, US Army Concepts Analysis Agency, CAA-TP-83-3, Bethesda, MD, August, 1983.

Shivaswamy, K. A. and Burns, P. J., 1993. *User's Manual for the Program CEMPLLOT, Version 1.0*, Internal Publication, Department of Mechanical Engineering, Colorado State University, Fort Collins, CO 80523, May, 1993

Stoll, George, 1992. *Personal Communication-PC Visualization Code*, US Army Concepts Analysis Agency, Bethesda, MD.

# THEORY OF A BIFURCATING MODEL NEURON: A NONLINEAR DYNAMIC SYSTEMS APPROACH\*

N.H. Farhat, M. Eldefrawy and S-Y. Lin  
University of Pennsylvania  
Electrical Engineering Department  
Philadelphia, PA 19104

**ABSTRACT.** A nonlinear dynamic systems approach to mathematical analysis of the integrate-and-fire (I & F) model neuron, a monoionic simplification of the Hodgkin-Huxley model for action potential generation in the excitable biological membrane, is presented. It is shown that under periodic activation (driving signal), as would arise in a synchronized network of such neurons when dendritic-tree processing is assumed, the firing behavior is described by an iterative map of the interval  $[0, 2\pi]$  onto itself which we call *phase-transition map* (PTM). Like other maps of the interval onto itself, the PTM can be studied employing the tools of nonlinear dynamics. This furnishes a novel way of viewing the microneurodynamics of neural networks and shows, that despite the simplifications made in the I & F model neuron, it exhibits, in its spiking behavior, a high degree of functional complexity approaching that of the living neuron. This is manifested by a variety of firing modalities depending on parameters of the periodic activation, which include regular firing phase-locked to the periodic activation or a subharmonic of it, quasi-periodic firing, erratic firing, and bursting and can bifurcate (rapidly switch) between these firing modalities as the parameters of the periodic activation are altered, hence the name *bifurcating neuron*. Illustrative examples of this complex behavior are given in the form of *bifurcation diagram*, *Arnold Tongues diagram* and the *Devil's Staircase diagram*. These show the bifurcating neuron is able to detect coherent episodes in its *incident spike wavefront*, the aggregate of spike trains incident on its synaptic inputs, and encodes such coherent activity whenever it occurs, in a complex manner depending on the parameters of the periodic activation potential produced by dendritic-tree processing of the incident spike wavefront. When the activation potential is not periodic, the bifurcating neuron reverts to the usual sigmoidal response and shows an upper limit on firing frequency which serves the useful function of containing the maximum firing activity in a network of such neurons in a manner analogous, but not exactly equivalent, to a similar limit imposed by refractoriness in the living neuron.

The bifurcating neuron model combines functional complexity with structural simplicity and low power consumption (because of its spiking nature). Accordingly, it is ideal for use in modeling, simulation, or hardware implementation of a new generation of neurocomputers in which synchronicity, bifurcation, and chaos, which are believed to underlie higher-level cortical functions, can be studied.

**INTRODUCTION.** Most neural net models assume sigmoidal neurons and therefore cannot account for the relative timing of action potentials (spike patterns) impinging on the dendritic-tree of a neuron from its presynaptic neurons. The bifurcating neuron concept and model [1] is an outcome of using the tools of nonlinear dynamics to characterize the behavior of the neuron's excitable (axonal) membrane in a manner that accounts naturally for temporal effects. It

---

\* Supported in part by the U.S. Army Research Office

combines functional complexity paralleling that of the living neuron with structural simplicity and low power consumption which makes it an attractive building-block for a new generation of neural networks specially suited for modeling higher-level cortical functions such as feature-binding, cognition, inferencing, attention, reasoning, etc.

There is increasing evidence in the literature encouraging speculation that such higher-level functions may involve synchronicity, phase-locking, bifurcations, erratic (chaotic firing for possible adaptive annealing), and dynamic signal dependent (adaptive) partitioning through which populations of cortical neurons may fleetingly divide into sub-populations or assemblies, each with a prescribed relative-phase pattern between its spiking neurons, that act in parallel in solving complex tasks with each sub-population acting collectively. Thus dynamic adaptive partitioning is the means by which cortical networks may carry out collective computations in several sub-populations of neurons in parallel. The significance of collective/parallel processing was recently discussed by Zak [2] in the context of *Nonlipschitzian* neural networks with unpredictable dynamics.

The relative timing and synchronicity in firing of neurons as a mechanism for feature-binding was predicted by Marlsburg [3] and Abeles [4] on theoretical grounds. Recently, synchronization effects in the firing of neurons at different recording locations in the cat's visual cortex upon suitable visual stimulus have been observed [5]-[11], raising thereby intriguing speculation about the nature of cortical processing, and stimulating wide spread interest. Temporal effects in the olfactory bulb have also been studied extensively by Freeman [12] and claims regarding the implications of rhythmic firing activity in consciousness have been covered by the media [13]. Further discussion of synchronization effects in temporal neural networks is given in [14]-[16]. It is now conjectured that synchronized firing of neurons in the visual cortex, for example, might : (a) label spatially coherent or common features in the visual data such as motion, contrast, texture, or color, (b) play a role in feature-binding, cognition, and other higher-level cortical functions, (c) could play a role in information absorption (learning) optimization in biological systems. Obviously, judging the validity of this conjecture requires thorough understanding of the mechanisms that can cause neurons at separate locations to synchronize their firings including, computer simulations aimed at studying and understanding the collective dynamics of spiking neural networks whose neurons are not simple processing element but possess functional complexity, like that of the bifurcating neuron model, that would reflect itself ultimately in the computing power of the network as a whole.

In this paper we present the results of a formulation of a bifurcating neuron theory which enables applying the power and the tools of nonlinear dynamics to characterizing its behavior in a manner that permits greater insight in neurodynamics and to possible utilization of synchronization, bifurcation and chaos in the design of a new generation of powerful neurocomputers.

The goal is to develop a bifurcating neuron theory, that is descriptive, predictive and qualitative. Descriptive in the sense that it should give true physical insight into the complicated processes involved. In the words of P.M. Koch [17], "a useful theory describes the essential physics of a process simply, preferably with figures and simple equations whose behavior with variation of parameters can be explored. The theory must be predictive because unless it can raise new questions and predict answers, it will likely be a dead end". Experimenters warm-up to theories that say "if you do such-and-such, then you will observe thus-and-so". The theory needs to be also quantitative because something is missing if it cannot be reduced to calculations that "get the numbers right".

All these goals are achieved in the theory presented here. To them we need to add, from the outset, the essential features we seek in a neuron model evolving from any developed theory. These features are:

- Production of an action potential or spike
- Existence of a refractory period to limit the maximum firing frequency to conserve energy and eliminate reverberations and reflections in a network
- Ability to integrate incoming synaptic inputs and to account for passive and/or active spatio-temporal processing carried out by the neuron's dendritic-tree
- Ability to respond to external signals in a manner that resembles that of the living (biological) neuron, i.e., production of spiking activity similar to that observed in the biological neuron under different forms of stimulus.
- Be structurally simple and possess low power consumption in order to facilitate hardware implementation of neurocomputing structures.

The starting point is the Hodgkin and Huxley model of the biological membrane [18] which describes membrane dynamics and spiking, the production of action potentials, in terms of three ionic currents. The shape of the action potential (spike) produced by this model resembles faithfully that observed in the living (biological) neuron and exhibits refractoriness. The distinct shape of the action potential is determined by the interplay between the dynamics of the three ionic currents in the model. Because the shape of all action potentials in biological networks is more or less the same, one can argue that the shape does not convey information and that only the interspike interval (the interval between action potentials) and/or the relative timing between action potentials in a network are important in neuroprocessing of information. This argument has prompted us to simplify the H-H model to a monoionic current model in order to facilitate its analysis as a dynamical system. This is done in Section 2 for two types of activation: nonperiodic, which is shown to lead to the usual sigmoidal response and periodic, which is assumed to arise when the dendritic-tree receives correlated spike trains, i.e., coherent spike wavefronts and is shown to lead to complex functional modalities far exceeding that of the sigmoidal neuron model. In Section 2 we also present examples of applying the bifurcating neuron theory, specifically the PTM, to characterizing the behavior of two bifurcating neuron embodiments that illustrate the complex behavior of the model. Concluding remarks and implications of such functional complexity and other properties of the bifurcating neuron are discussed in Section 4.

**2. ANALYSIS.** Figure 1(a) shows an equivalent circuit of the monoionic model. In it  $I$  is the monoionic membrane current representing an energy source for restoring the membrane potential after firing,  $R$  and  $C$  are the membrane resistance and capacitance respectively,  $v$  is the capacitor voltage,  $i = \phi(v')$  represents the membrane nonlinearity, assumed to be S-shaped and  $u(t)$  is the activation potential of the membrane.  $u(t)$  represents the effect produced at the biological neurons hillock by synaptic inputs to the neuron. A piece-wise linear approximation of the S-shaped nonlinearity is shown in Fig. 2 after inclusion of the effect of the activation potential  $u(t)$  which represents the modulation in membrane potential produced by integration of action potentials (input spike trains) arriving at the neuron's dendritic-tree from its presynaptic neurons. The circuit in Fig. 1(b) is equivalent to that in Fig. 1(a) when the voltage source  $E$  is set equal to  $E = IR$ . It represents the circuit diagram of an integrate-and-fire neuron or relaxation oscillator neuron whose dynamics we will study here in detail. Despite its simplicity, this circuit will be shown to exhibit complex behavior, specially when  $u(t)$  is periodic. In Fig. 2,  $v_{th}$  and  $v_{ext}$  are related respectively to the breakdown or threshold voltage  $v_1$  and the extinction voltage  $v_2$  of the S-shaped nonlinearity  $\phi(v)$ . Notice that in the absence of activation potential or (driving signal)  $u(t)$ ,  $v'$  coincides with  $v$  and  $v_{th}$  reduces to  $v_1$  while  $v_{ext}$  reduces to  $v_2$ .

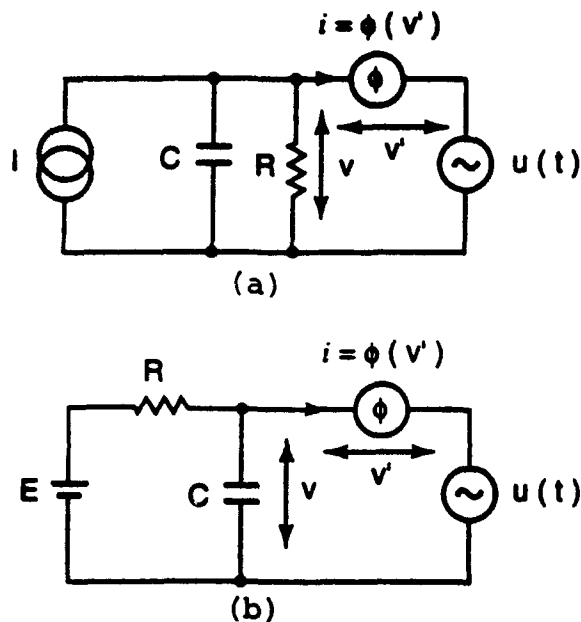


Fig. 1. Equivalent circuit of monoionic model of the excitable membrane.

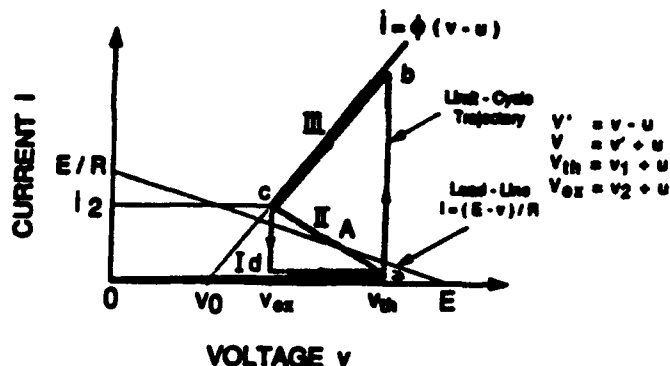


Fig. 2. Piece-wise linear approximation of the S-shaped nonlinearity of the neurons membrane.

In the absence of the driving signal  $u(t)$  and when  $E > v_1$ , the circuit of Fig. 1(b) behaves as a nonlinear relaxation oscillator. The S-shaped  $i-v$  (current-voltage) characteristic  $i = \phi(v')$  shown in Fig. 2 reduces to that shown in Fig. 3, i.e.,  $i = \phi(v)$  curve with a negative resistance region, a breakdown or threshold voltage  $v_1$  and extinction voltage  $v_2$ . This  $i-v$  characteristic  $i = \phi(v)$  is seen to consist of three distinct regions I, II and III. Regions I and III are positive resistance regions corresponding to the resistance of nonlinear element  $i = \phi(v)$  in the off and on states respectively, while region II is a negative resistance region. The distance  $\overline{cd}$  in this plot is exaggerated in comparison to that in actual  $i-v$  characteristics in order to delineate the negative resistance region of the plot. The behavior of the circuit in Fig. 1(b) is governed by the nonlinear differential equation,

$$(1) \quad C \frac{dv}{dt} + \phi(v) = \frac{E - v}{R}$$

The steady state,  $\frac{dv}{dt} = 0$ , of eq. (1) is defined by

$$(2) \quad \phi(v) = \frac{E - v}{R}$$

in which  $i = (E-v)/R$  defines a load-line in the  $i-v$  plane as shown in Fig. 3 and the intersection point between the load-line and  $\phi(v)$  defines the operating point A.

When  $E > v_1$  and the load-line  $i(v) = (E-v)/R$  intersects the  $i-v$  characteristics in the negative resistance region i.e., the operating point A falls in region II as shown in Fig. 3, the circuit of Fig. 1(b) exhibits the *limit-cycle* oscillations indicated by the closed dotted line trajectory  $\overline{abcd}a$ . The portion  $\overline{abcd}$  of the trajectory corresponds to the conduction of a current spike through the nonlinear element when the capacitor voltage  $v(t)$  reaches the breakdown value  $v_1$ , while the portion  $\overline{da}$  corresponds to the interspike interval during which the charge in capacitor C is being restored and the nonlinear element is in off state. Changes in the values of E, v, and/or R, cause the load-line and operating point A to shift. Shifting the location of A within the negative resistance regions alters the speed of motion along the limit-cycle trajectory and this alters the firing frequency. Limit-cycle oscillations cease to exist when the operating point A enters either of regions I or III. When A is in region III, the nonlinear element remains on while for an A falling in region I, the nonlinear element remains off. The simple circuit dynamics described qualitatively above become considerably more complicated in the presence of a time varying or periodic driving signal  $u(t)$ . Then we need to modify eq. (1) by replacing  $\phi(v)$  with  $\phi(v - u(t))$  which means that the limit-cycle dynamics are complicated by time dependent displacement of the  $\phi$  curve in the horizontal direction vis-a-vis the stationary load line. This and the interplay between  $u(t)$  and the capacitor voltage  $v(t)$  is the underlying cause of the complex dynamics of the bifurcating neuron circuit and the observation of complex firing sequences for certain values of parameters of the periodic driving signal  $u(t)$  as will be shown later.

The limit-cycle trajectory can be readily displayed on a CRO by driving the x-axis of the CRO with the voltage drop proportional to  $i(t)$  in Fig. 3 and the Y-axis with the voltage  $v = v(t)$  appearing across the nonlinear element. An example of such a display is given in the photograph of Fig. 4 which represents the *limit-cycle trajectory* in the  $i-v$  phase-space of the circuit of Fig. 1(b) and is seen to be similar to the idealized limit-cycle trajectory marked in Fig. 3. The apparent uneven brightness of the trajectory in the photograph, reflects the uneven speed with which the electron beam traces the trajectory on the CRO's phosphor screen as dictated by the time variation of  $i(t)$  and  $v(t)$ .

The current waveform  $i(t)$  corresponding to limit-cycle oscillations consists of a train of narrow fixed-shape impulses, or spikes, of fixed peak amplitude  $i_0$  and duration  $T_1$  corresponding to the  $a \rightarrow b$ ,  $b \rightarrow c \rightarrow d$  portions of the trajectory. The interspike interval  $T_2$  corresponds to the  $d \rightarrow a$  interval of the trajectory. The spikes in  $i(t)$  represent *action potentials* so to speak, of the bifurcating neuron. Expressions for  $T_1$  and  $T_2$  can be readily derived when the nonlinear element in Fig. 1 is assumed to have the idealized S-shaped characteristic of Fig. 3. The results are given by [19],

$$(3) \quad T_1 = RC \ln \left( \frac{E - v_2}{E - v_1} \right)$$

$$(4) \quad T_2 = \rho C \ln \left( \frac{(v_1 - v_2)R - (E - v_1)R_i}{(v_2 - v_0)R - (E - v_2)R_i} \right)$$

where  $\rho = RR_i/(R + R_i)$ , with  $R_i$  being the nonlinear element's resistance in segment III of the  $i-v$  characteristics, and  $v_0$  is the voltage value at the point of intersection of the extension of segment III of the characteristic with the  $v$  axis as shown in Fig. 3. All other quantities in eqs. (3) and (4)

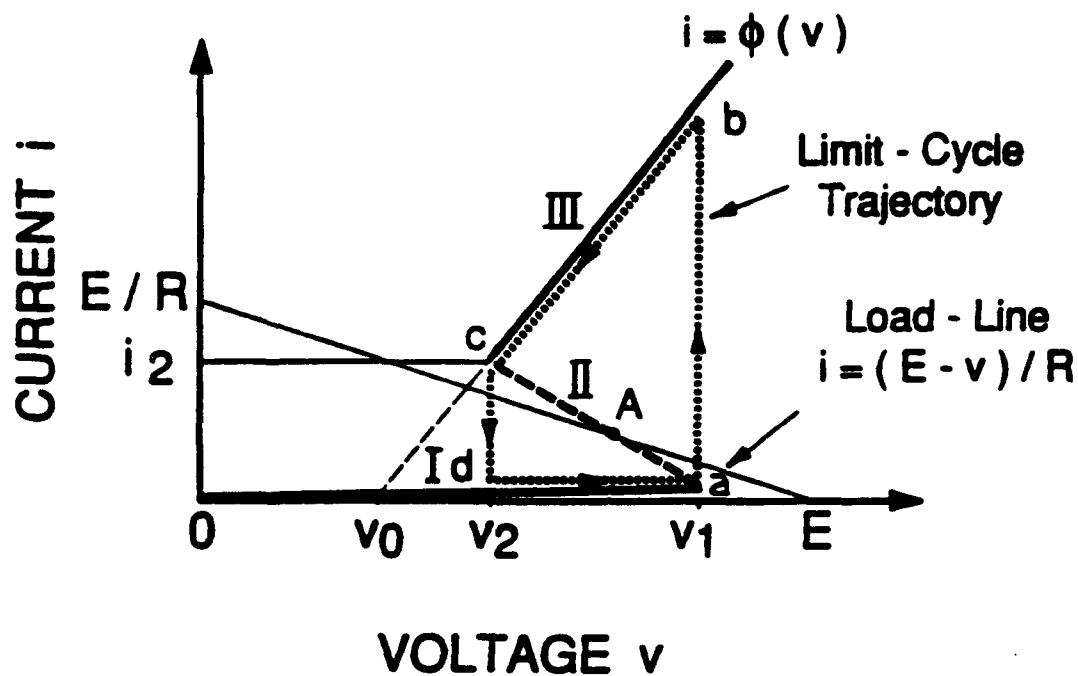


Fig. 3. Piece-wise linear approximation of the  $i$ - $v$  (current-voltage) characteristic of the glow-lamp (glow-discharge tube) consisting of three segments: I, II and III and showing the limit-cycle trajectory  $abcd$  (dotted line) referred to in the text.

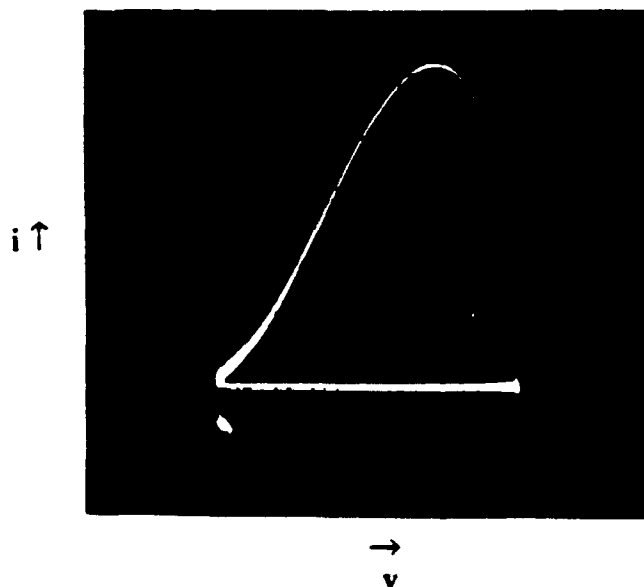


Fig. 4. Limit cycle trajectory in the  $i$ - $v$  phase-space of the circuit of Fig. 1 associated with the natural oscillations of the bifurcating neuron ( $E = 160$  [V],  $R = 100$  [K $\Omega$ ],  $C = .1$  [ $\mu$ F],  $R_s = 10$  [ $\Omega$ ]). Vertical scale: 10 [mA/div.], horizontal scale: 10 [V/div.]. A.C. coupled CRO display is shown.

are as defined earlier. Note that eqs. (3) and (4) are meaningful only when  $E > v_1$ . When this condition is not satisfied the arguments of the  $\ln(\dots)$  term in eq. (3) becomes negative precluding thereby physically meaningful solutions and disruption of limit-cycle oscillations. In accordance to eqs. (3) and (4) the interspike-interval or period of the bifurcating neuron's oscillation is  $T = T_1 + T_2$  and the instantaneous firing frequency is  $f = 1/T = 1/(T_1 + T_2)$ . Because, in normal operation,  $R_i \ll R$ , we find that the spike width  $T_2$  is much less than the interspike interval  $T_1$  and therefore  $T \approx T_1$  becomes a good approximation.

It is convenient to consider two regimes of operation when  $u(t) \neq 0$ . One is the none self-oscillatory regime, and the other is the self-oscillatory regime. We consider next the regime when the supply voltage  $E$  is slightly less than the breakdown voltage  $v_1$ , so that the nonlinear element is normally in extinguished state and the limit-cycle oscillations are not triggered spontaneously. The subthreshold value of  $E$  in this case is such that the addition of the driving signal is sufficient to trigger the limit-cycle oscillation whenever the voltage across the nonlinear element exceeds threshold. Thus, in the presence of  $u(t)$  and specially when  $u(t)$  is sufficiently slowly varying, that is the scale of its time-variations is large compared to the natural oscillation period  $T$  of the circuit we discussed earlier, the effect of  $u(t)$  on the circuit can be shown to be reproduced by suitably changing the supply voltage  $E$ . In this regime, our bifurcating neuron exhibits nearly sigmoidal dependence of firing frequency on activation i.e., on the effective voltage across the nonlinear element.

To show this we proceed as follows: Because the spike width  $T_2$  is very narrow, we approximate the spike portion  $abcd$  of the limit-cycle oscillation of the S-shaped nonlinearity in Fig. 3 or Fig. 2 by,

$$(5) \quad i = \phi(v) = \phi(v') = \begin{cases} 0 & v' < v_1 \\ i_0 \delta(t) & v \geq v_1 \end{cases}$$

Since,

$$v' = v - u$$

we have,

$$(6) \quad \begin{aligned} v_1 &= v_{th} - u \rightarrow v_{th} = v_1 + u \\ v_2 &= v_{ex} - u \rightarrow v_{ex} = v_2 + u. \end{aligned}$$

The spike amplitude  $i_0$  in eq. (5) is determined by considering the change  $\Delta Q$  in the charge stored in the capacitor  $C$  because of a single firing of the neuron, i.e., the discharge of the capacitor during the  $abcd$  portion of a limit-cycle oscillation. This is,



$$\Delta Q = C \Delta v$$

or,

$$\int i dt = C(v_{th} - v_{ex}) = C(v_1 - v_2)$$

or by using eq. (5)

$$\int i_o \delta(t) dt = C(v_1 - v_2)$$

$$(7) \quad i_o = C(v_1 - v_2).$$

which is seen to be determined by the capacitance  $C$  and the parameters  $v_1$  and  $v_2$  of the nonlinear S-shaped nonlinearity.

Now when  $v < v_{th}$ ,  $\phi(v - u) = 0$ , eq. (1) reduces to,

$$C \frac{dv}{dt} = \frac{E - v}{R}$$

whose solution is,

$$(8) \quad v(t) = E - (E - v_2 - v(o)) e^{-\frac{t}{RC}}$$

which represents the capacitor voltage build-up in the  $d \rightarrow$  a region of the limit-cycle trajectory in

Fig. 2.

The interspike interval  $T$  is found from eq. (8) by letting  $t = T$  and  $v(T) = v_{th}(T) = v_1 + u(T)$ . This yields,

$$(9) \quad T = RC \ln \frac{E - v_2 - u(o)}{E - v_1 - u(T)} = RC \ln \frac{E - v_{ex}(o)}{E - v_{th}(T)}$$

This analytical model is known as the *integrate and fire* model of the spiking neuron.

We see from comparing eqs. (9) and (6) that the effect of the activation potential (the neuron's driving signal)  $u$  is to horizontally shift the curve  $\phi(v - u)$  in Fig. 2 to the left or to the right by an amount that depends on the magnitude and sign of  $u$ . Because the load line is stationary, this horizontal shifting causes a migration of the operating point A on the  $\phi$  curve.

When  $u$  is positive and is increasing,  $\phi(v - u)$  would be shifted to the right until  $v_{th} \geq E$  at which instant the operating point A enters segment I of the  $\phi$  curve and limit-cycle oscillations are halted or extinguished in agreement with the prediction of eq. (9) when  $v_{th} \geq E$ . This is analogous to complete inhibition of the firing of a biological neuron when a sufficient inhibitory activation potential  $u$  is present. Thus a positive  $u$  in our neuron model corresponds to inhibitory signal.

When  $u$  is negative and is gradually decreased,  $\phi(v - u)$  shifts gradually to the left. This has the effect of increasing the firing frequency. A negative  $u$  in our neuron circuit (Fig. 1(b)) is therefore equivalent to an excitatory input. As the value of  $u$  is made more negative (increased excitation) the shift of  $\phi(v')$  to the left continues until the operating point A leaves the negative resistance segment II and enters the positive resistance segment III of the  $\phi$  curve. When this happens, the nonlinear element  $\phi$  remains on (conducting state) and limit-cycle oscillations are again extinguished. Limit-cycle oscillations can not be triggered then no matter how strong an excitatory signal is received. This has the effect of limiting the maximum possible firing frequency and is analogous to the limiting of the maximum firing frequency in the biological neuron by the presence of an absolute refractory period.

We examine next two cases of  $u(t)$ . In one case  $u(t) = -V_m$  where  $V_m$  is a positive real constant whose value is changed gradually to determine how  $T$  and hence the firing frequency  $f = \frac{1}{T}$  changes. This case leads to conventional sigmoidal response with an upper limit on the firing frequency. In the second case  $u(t)$  is assumed to be periodic. It leads to complex firing modalities of the neuron and includes bifurcation between them depending on parameters of the periodic activation.

### Case I

Let  $u(t) = -V_m$ , then eq. (9) becomes,

$$(10) \quad T = RC \ln \frac{E - v_2 + V_m}{E - v_1 + V_m}$$

and  $\phi(v - u) = \phi(v + V_m)$ . Thus a positive increasing  $V_m$  corresponds to increased inhibition and a lowering of the firing frequency  $f = \frac{1}{T}$ , while a negative decreasing  $V_m$  corresponds to increased

excitation. The value of the maximum firing frequency  $f_{\max}$  and the value of  $V_m$  at which the maximum firing frequency occurs are determined from Fig. 2 in the following manner:

When  $\phi$  shifts to the left so that the load line passes through point C we have,

$$\frac{\bar{cd}}{E - V_{ex}} = \frac{E/R}{E}$$

or,

$$\frac{\bar{cd}}{E - v_2 - V_m} = \frac{1}{R}$$

which yields the maximum value of  $V_m$  for which firing stops,

$$(11) \quad V_{m \max} = E - v_2 - \bar{cd}R$$

by combining eqs. (8) and (7) one obtains an expression for the maximum firing frequency of our neuron.

$$(12) \quad f_{\max} = \frac{1}{T_{\min}}$$

where

$$(13) \quad T_{\min} = RC \ln \frac{2E - 2v_2 - \bar{cd}R}{2E - (v_1 + v_2) - \bar{cd}R}$$

Thus the circuit in Fig. 1(b) is seen to have, when  $u(t) = -V_m$ , an upper limit on the firing frequency determined by the value of  $V_m$ . The upper bound on firing frequency is imposed by the nonlinear element remaining in the on state which stops limit-cycle oscillations. In contrast the maximum firing frequency of a biological neuron is imposed by the absolute refractory period immediately following each action potential (spike firing) during which the neuron is incapable of firing again no matter how strong an excitatory stimulus it receives. In limiting the maximum firing frequency, refractoriness in the living neuron helps conserve energy and eliminate reverberations.

The dependence of firing frequency  $f = \frac{1}{T}$  on  $V_m$  where  $T$  is given by eq. (10) is shown in Fig. 5 (solid line curve marked  $T_r = 0$ ) together with the measured firing frequency (dotted line curve). The two curves are for the case when the S-shaped nonlinearity  $\phi$  is that of a glow-lamp. The circuit parameters and glow-lamp parameters are given in the figure caption. Verification of eq. (12) for an example of glow-lamp S-shaped nonlinearity with following parameters:

$$\begin{array}{lll} E = 160 \text{ [V]} & , & v_1 = 141 \text{ [V]} & , & v_2 = 125 \text{ [V]} \\ R = 10^5 \text{ [\Omega]} & , & C = 10^{-7} \text{ [F]} & , & \bar{c}d = 4 \cdot 10^{-4} \text{ [A]} \end{array}$$

yields,

$$T_{\min} = 7.6 \cdot 10^{-3} \text{ [sec]}$$

and therefore

$$f_{\max} = \frac{1}{T_{\min}} = 131 \text{ [Hz]}$$

which is in agreement with the experimental cutoff frequency observed in the preceding figure. Figure 5 also shows the effect of arbitrarily adding an absolute refractory interval  $T_r$  [msec] to the right-hand side of eq. (10). Such ad hoc inclusion of  $T_r$  is seen to accelerate saturation of the firing frequency.

## Case II

The preceding analysis shows basically the bifurcating model neuron can exhibit usual sigmoidal response. We show next however, that when the activation potential  $u(t)$  is periodic, the neuron alters its behavior and is able to phase-lock its firing to the frequency of the periodic activation or a sub-harmonic of it, or can fire quasiperiodically, erratically, or in bursts, all depending on the amplitude and frequency of the activation potential. Periodic activation at the neuron's hillock is assumed to arise whenever the spike trains (action potentials) incident on the neuron's dendritic-tree are correlated. If we refer to the aggregate of all spike trains incident, *at any time*, on the dendritic-tree as the *incident spike wavefront*, then a coherent incident spike wavefront produces a periodic activation potential, i.e., a periodic driving signal for the neuron. Thus we examine now the behavior of the circuit of Fig. 1(b) when the activation potential  $u(t)$  is periodic. The main result of the analysis is a Phase-Transition Map (PTM) which relates the phase  $\theta_{n+1}$  of the  $n+1$  spike produced by the neuron to  $\theta_n$ , the phase of the  $n$ -th (preceding) spike. In our formulation, the phase of a spike is always measured relative to the immediately preceding peak (or some other selected feature) of the periodic activation potential. The PTM is a nonlinear iterative map on the  $(0 - 2\pi)$  interval, and as such, it lends itself to further analysis and treatment as is usually done in nonlinear dynamics with other maps of the interval onto itself like the *logistic* and the *circle map* for example.

For simplicity we assume the periodic activation or driving signal of the neuron (essentially the membrane potential at the hillock) is cosinusoidal of amplitude  $a$ , radian frequency  $\omega_s$  and fixed phase  $\theta_0$ , i.e.,

$$u(t) = a \cos(\omega_s t + \theta_0)$$

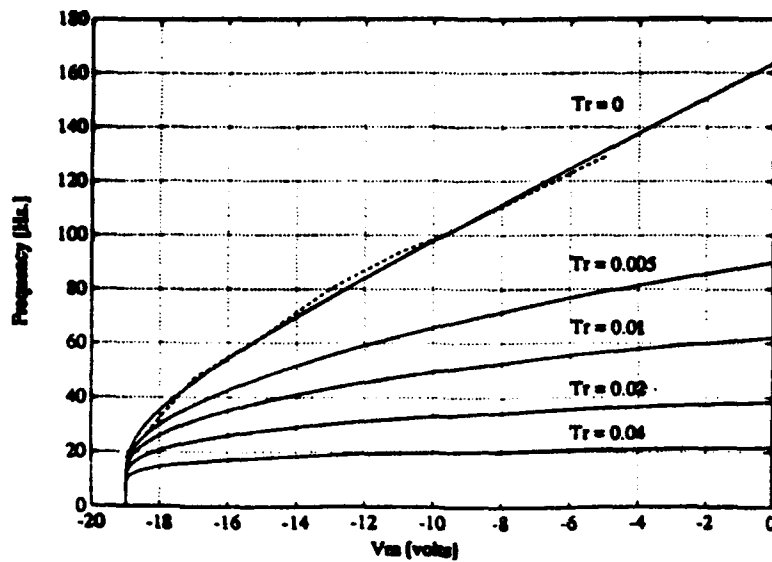


Fig. 5. Dependence of Firing frequency on activation potential  $V_m$ . Example of glow-lamp S-shaped nonlinearity with the following fixed parameters:  $V_1 = 141$  [volts]  $V_2 = 125$  [volts],  $RC = 0.01$  [sec.],  $E = 160$  [volts]. (solid) computed, (---) measured.

Then starting from the expression for the capacitor voltage in Fig. 1(b) we derived earlier (eq. 8),

$$(14) \quad v(t) = E - (E - v_2 - u(0)) e^{-\frac{t}{RC}}$$

and by referring to Fig. 6 we see that

$$(15) \quad v(t) = v_{th}(t) \quad \text{at} \quad t = T$$

where

$$(16) \quad v_{th}(t) = v_1 + a \cos(\omega_s t + \theta_0)$$

Therefore from eqs. (14) and (15),

$$(17) \quad E - v_1 - a \cos(\omega_s T + \theta_0) = (E - v_2 - a \cos \theta) e^{-\frac{T}{RC}}$$

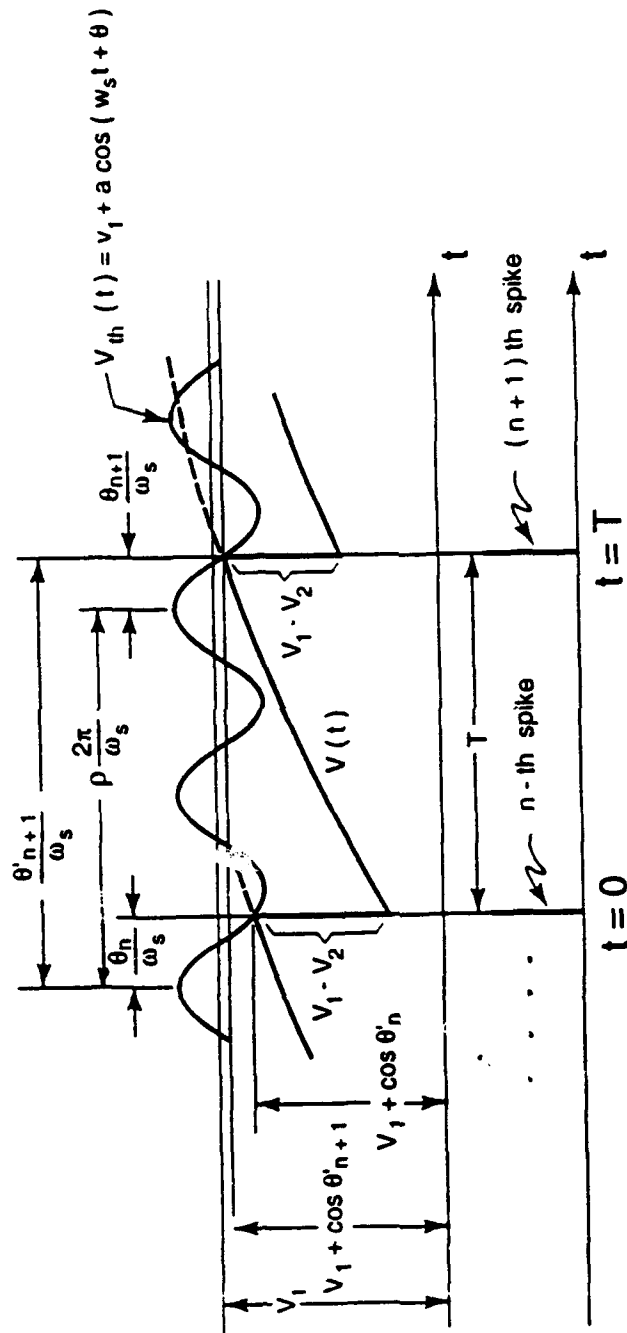


Fig. 1. Relation between the periodic driving signal (activation potential) appearing in  $V_{th}(t)$ , the capacitor voltage  $V(t)$ , and the spike train produced by the neuron (only two spikes are shown) referred to in the accompanying analysis. Note the relation between  $\theta_n$ ,  $\theta'_{n+1}$  and  $\theta_{n+1}$ .

$$(18) \quad \omega_s T + \theta_n = \theta'_{n+1}$$

or,

$$(19) \quad \omega_s T = \theta'_{n+1} - \theta_n$$

$$(20) \quad T = \frac{\theta'_{n+1} - \theta_n}{\omega_s}$$

and letting

$$(21) \quad \theta'_{n+1} = \theta_{n+1} + m2\pi \quad m = \text{integer}$$

we have,

$$(22) \quad \theta_{n+1} = \theta'_{n+1} - m2\pi = [\theta'_{n+1}] \bmod 2\pi$$

Therefore from eqs. (14) and (17),

$$(23) \quad E - v_1 - a \cos \theta'_{n+1} = (E - v_2 - a \cos \theta_n) e^{-\frac{\theta'_{n+1} - \theta_n}{\omega_s RC}}$$

or, in terms of normalized variables,

$$(24) \quad (1 - v'_1 - a' \cos \theta'_{n+1}) e^{\frac{\theta'_{n+1}}{x}} = (1 - v'_2 - a' \cos \theta_n) e^{\frac{\theta_n}{x}}$$

where,

$$v'_1 = v_1/E, \quad v'_2 = v_2/E, \quad a' = a/E, \quad x = \omega_s RC$$

Therefore finally from (24) and (22),

$$(25) \quad \theta'_{n+1} = \theta_n + x \ln \left( \frac{1 - v'_2 - a' \cos \theta_n}{1 - v'_1 - a' \cos \theta'_{n+1}} \right)$$

and

$$(26) \quad \theta_{n+1} = [\theta'_{n+1}]_{\text{mod.} 2\pi}$$

Equations (25) and (26) are the main results of this analysis. Taken together they form the relationship or mapping between  $\theta_n$ , the phase of the  $n$ -th spike and  $\theta_{n+1}$ , the phase of the  $(n+1)$ th spike in the firing activity of the neuron under cosinusoidal activation. Note the phase of a spike is always measured from the immediately preceding peak of the cosinusoidal drive signal (see Fig. 6). Equations (25) and (26) can be expressed symbolically in the form

$$(27) \quad \theta_{n+1} = g(\theta_n)$$

where the function  $g(\cdot)$  is defined by the mapping in eqs. (25) and (26). Note that eq. (25) is a transcendental equation in  $\theta_{n+1}$  that must be solved first given  $\theta_n$ , the system parameters, and those of the periodic driving signal. Having obtained  $\theta'_{n+1}$  its modulus  $2\pi$  is computed to obtain  $\theta_{n+1}$ . We call the mapping in eq. (27) *Phase-Transition Map (PTM)*. The PTM is a nonlinear iterative map of the interval  $[0, 2\pi]$  onto itself. Like other iterative maps of interval onto itself, such as the Logistic Map and the Circle Map, the PTM can be studied using the tools of nonlinear dynamical systems.

For example, Fig. 7 shows the steps involved in obtaining a plot for the PTM of the bifurcating neuron using eq. (25) assuming a glow-lamp S-shaped nonlinearity and a periodic cosinusoidal driving signal activation of amplitude  $a = 1$  [V] and frequency  $f_s = \frac{\omega s}{2\pi} = 190$  [Hz]. The PTM for this case, shown in Fig. 7 (c), can be iterated graphically as illustrated in Fig. 8(a). Entering the abscissa of this figure from an initial value  $\theta_0$  one draws a vertical line that intersects the plot giving the value of  $\theta_1$  which can be re-entered on the abscissa and the process repeated to find  $\theta_2$  and so forth. This procedure is greatly simplified by using the  $45^\circ$  line  $\theta_{n+1} = \theta_n$  in performing the iterations. From  $\theta_0$  one moves vertically to intersect the PTM plot, then moves horizontally to intersect the  $45^\circ$  line, at a point whose abscissa gives the value of  $\theta_1$ , then move



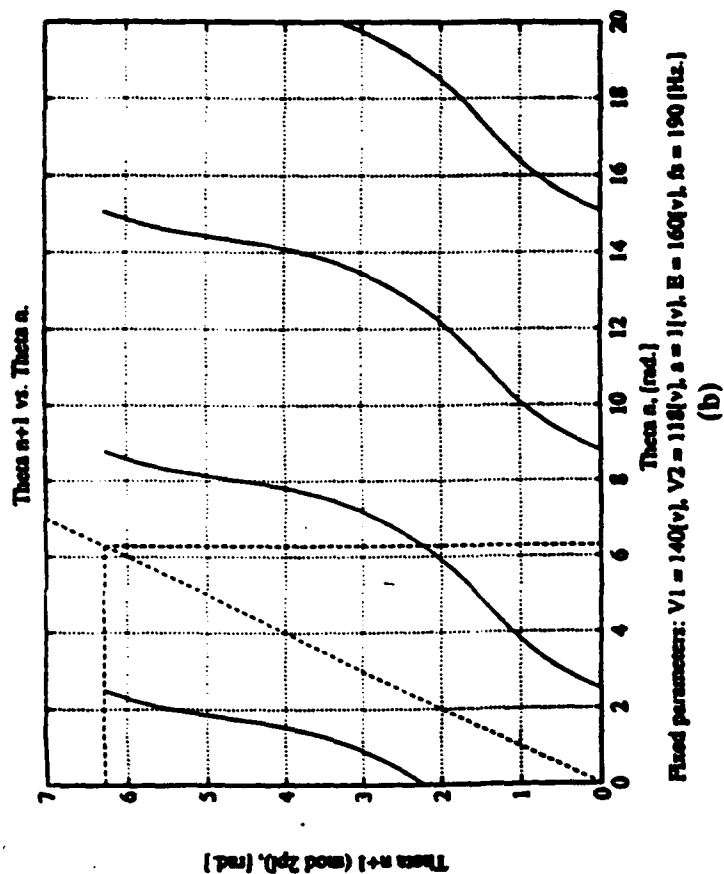
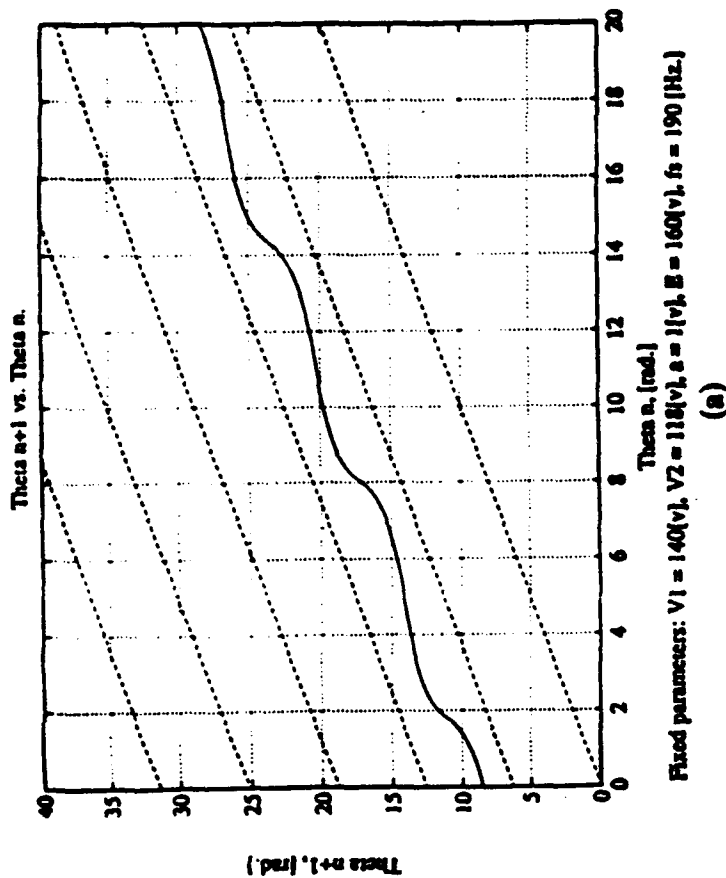
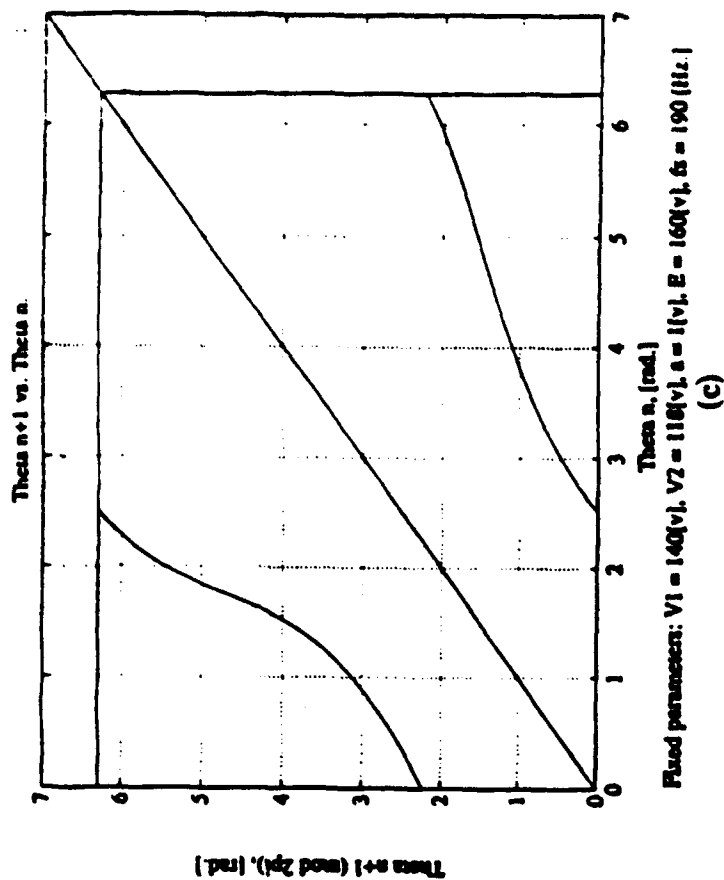


Fig. 7

STEPS IN DEVELOPING THE PHASE TRANSITION MAP FOR THE SINUSOIDALLY DRIVEN ( $a = 1$  [v],  $f_s = 190$  [Hz]) BIFURCATING NEURON ASSUMING GLOW-LAMP S-SHAPED NONLINEARITY. (a)  $\theta_{n+1}$  vs.  $\theta_n$  FROM eq. 25, (b)  $[\theta_{n+1}] \bmod 2\pi$  vs.  $\theta_n$ , (c)  $[\theta_{n+1}] \bmod 2\pi$  vs.  $[\theta_n] \bmod 2\pi$ .



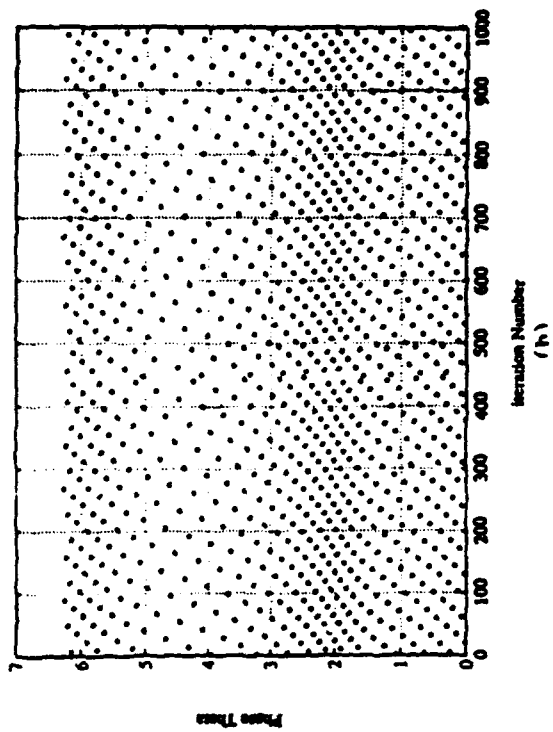
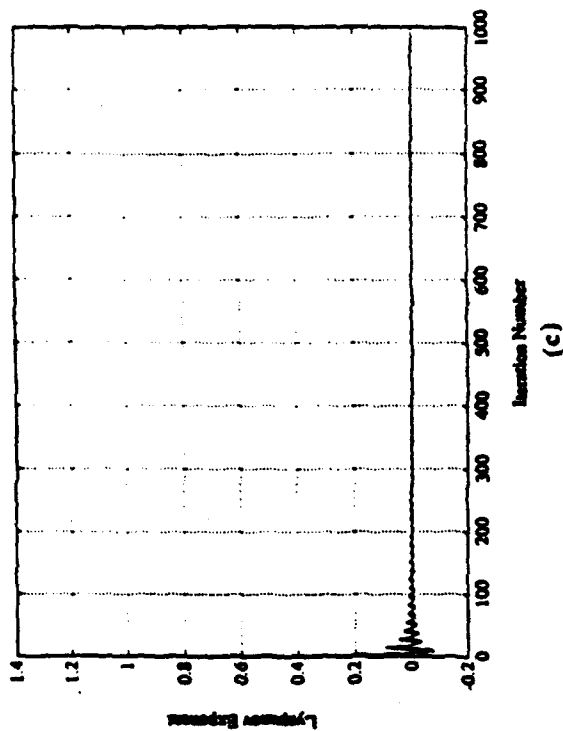
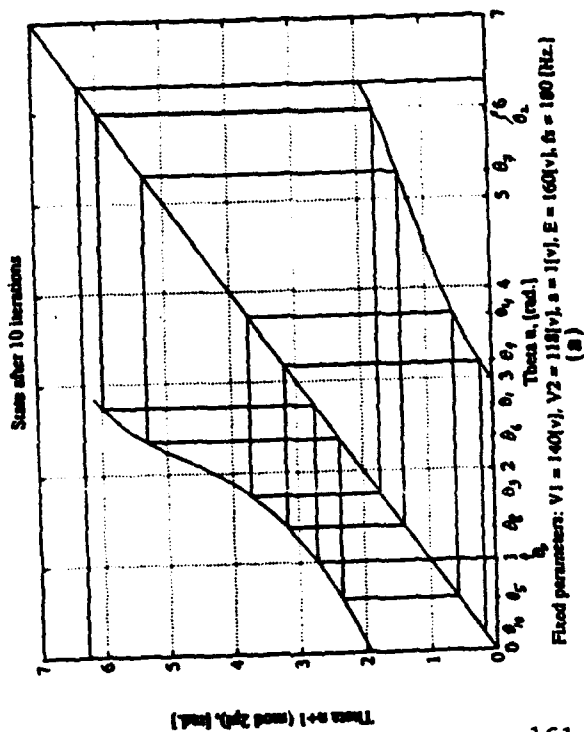


Fig. 8  
RESULTS OF NUMERICAL ITERATION OF THE  
PHASE-TRANSITION MAP OF THE CO-  
SINUSOIDALLY DRIVEN ( $a = 1(\text{V})$ ,  $f_s = 190(\text{Hz})$ )  
BIFURCATING NEURON: (a) FIRST 1000  
ITERATIONS. (b)  $\theta_n$  vs. ITERATION NUMBER  $n$ ,  
(c) LYAPUNOV EXPONENT.

vertically to intersect the plot again and move from there horizontally to meet the 45° line and obtain  $\theta_2$  and so forth. Starting from any initial value  $\theta_0$  the resulting sequence of phase values  $\theta_n$   $n=1,2,3...$  eventually settles into a regular pattern (orbit) or seemingly an erratic one depending on the parameters of the plot. The values of  $\theta_n$  vs.  $n$  where  $n$  is the iteration number is shown in Fig. 7(b) for 1000 iterations. This plot clearly shows the firing pattern of the neuron for this case is quite complex but regular. Figure 8(c) shows the Lyapunov exponent [20],

$$(28) \quad \lambda = \lim_{N \rightarrow \infty} \left[ \frac{1}{N} \sum_{n=1}^N |g'(\theta_n)| \right]$$

where  $g'(\theta_n)$  is the derivative or slope of the PTM at the iteration points, the values of  $\theta_n$  produced by the map, and the periodically driven bifurcating neuron it represents. The Lyapunov exponent is a measure of regularity, or lack of it, in iterative maps. A value of  $\lambda > 0$  is usually taken as an indication of irregularity or chaos. Because the  $\theta_n$  pattern in Fig. 8(b) is regular, the corresponding Lyapunov exponent is seen to stabilize to zero after few hundred iterations when transients die out. The shape of the PTM changes markedly when the parameters ( $a, f_s$ ) of the cosinusoidal driving signal are altered. This is demonstrated in the PTM plots and associated  $\theta_n$  orbits which show period three (left) and period six (right) firing modalities shown in Fig. 9. Again because the  $\theta_n$  patterns are ordered the Lyapunov exponents of the two plots are negative. It is worth noting that period-N firing modality covers the case when the neuron is bursting.

Qualitatively, similar results are obtained for bifurcating neurons employing solid-state nonlinear elements with S-shaped nonlinearity such as the unijunction transistor (UJT) and the programmable unijunction transistor (PUT) which are solid-state equivalents of the glow lamp.

A more encompassing view of bifurcating neuron dynamics, is provided by the *bifurcation diagram*. This is an intensity plot of the resulting phase orbit or sequence  $\theta_n$ , after transients are allowed to die down. The values of  $\theta_n$  are entered as points along the vertical above each frequency point as shown in the example of the measured bifurcation\* diagram for a programmable unijunction transistor neuron (PUTON) embodiment of the bifurcating neuron shown in Fig. 10 for a fixed driving signal amplitude  $a = 0.175$  [V]. Note the values of  $\theta_n$  in the figure are shown normalized to  $2\pi$ . The figure demonstrates clearly, and at one glance, the rich variety of firing modalities the PUTON goes through as the frequency of periodic activation (driving signal) is altered.

---

\*Circuit diagram of the PUTON and the  $\theta_n$  measurement system are not given here but will be included in a future publication.

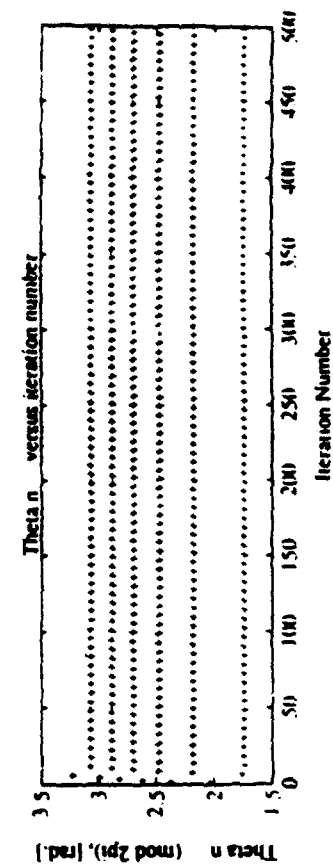
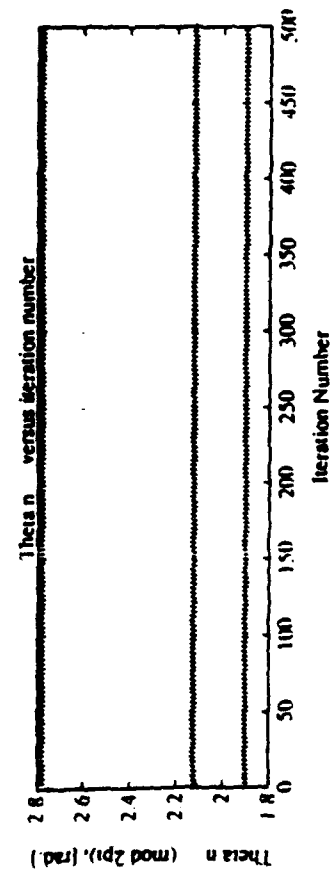
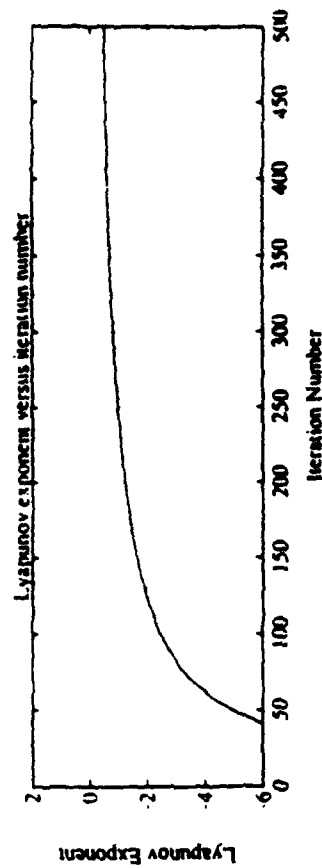
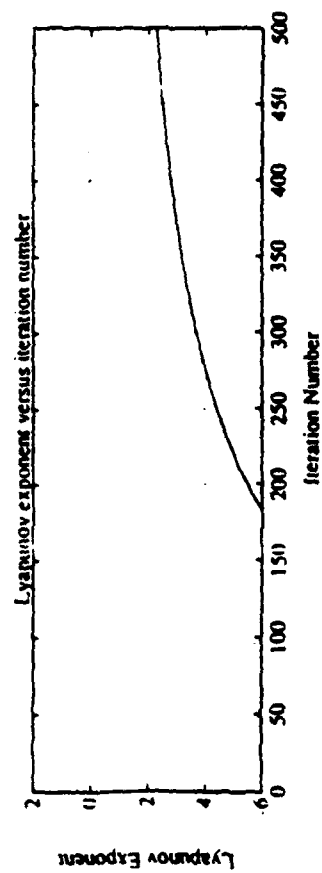
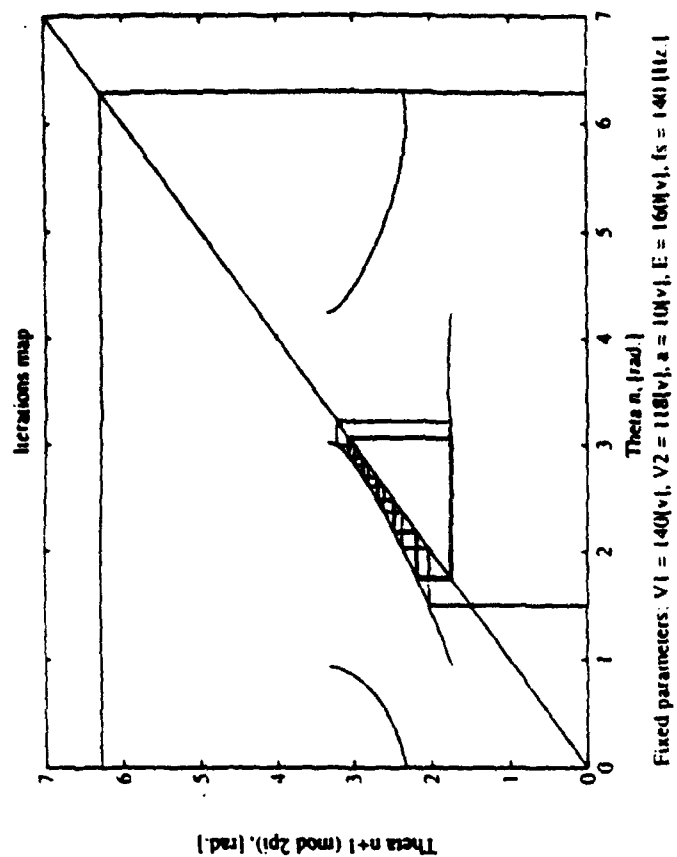
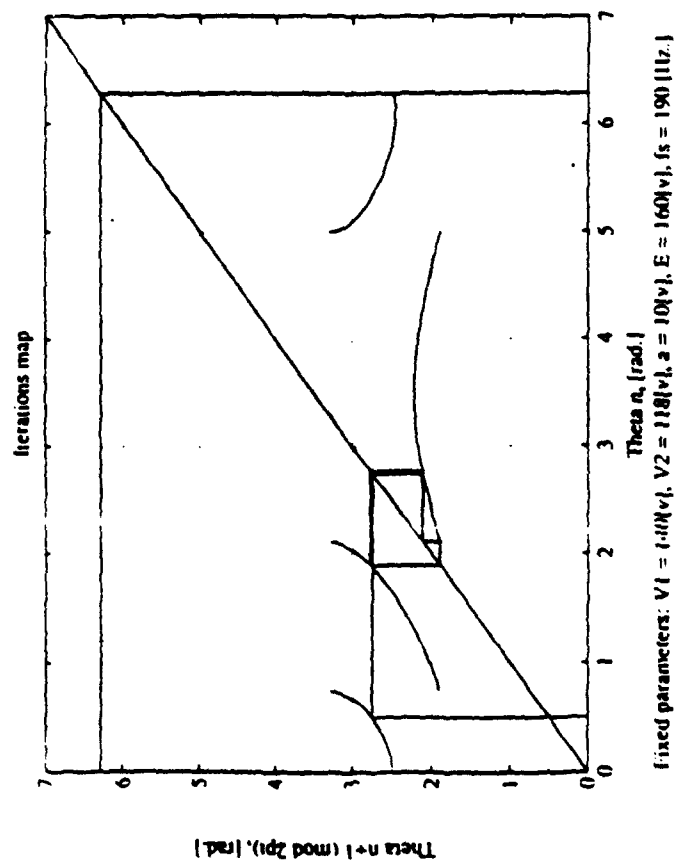


Fig. 9. Two examples illustrating how the PTM, the resulting  $\Theta_n$  sequence, and the associated Lyapunov exponent are altered by altering the driving signal parameter  $f_s$ .

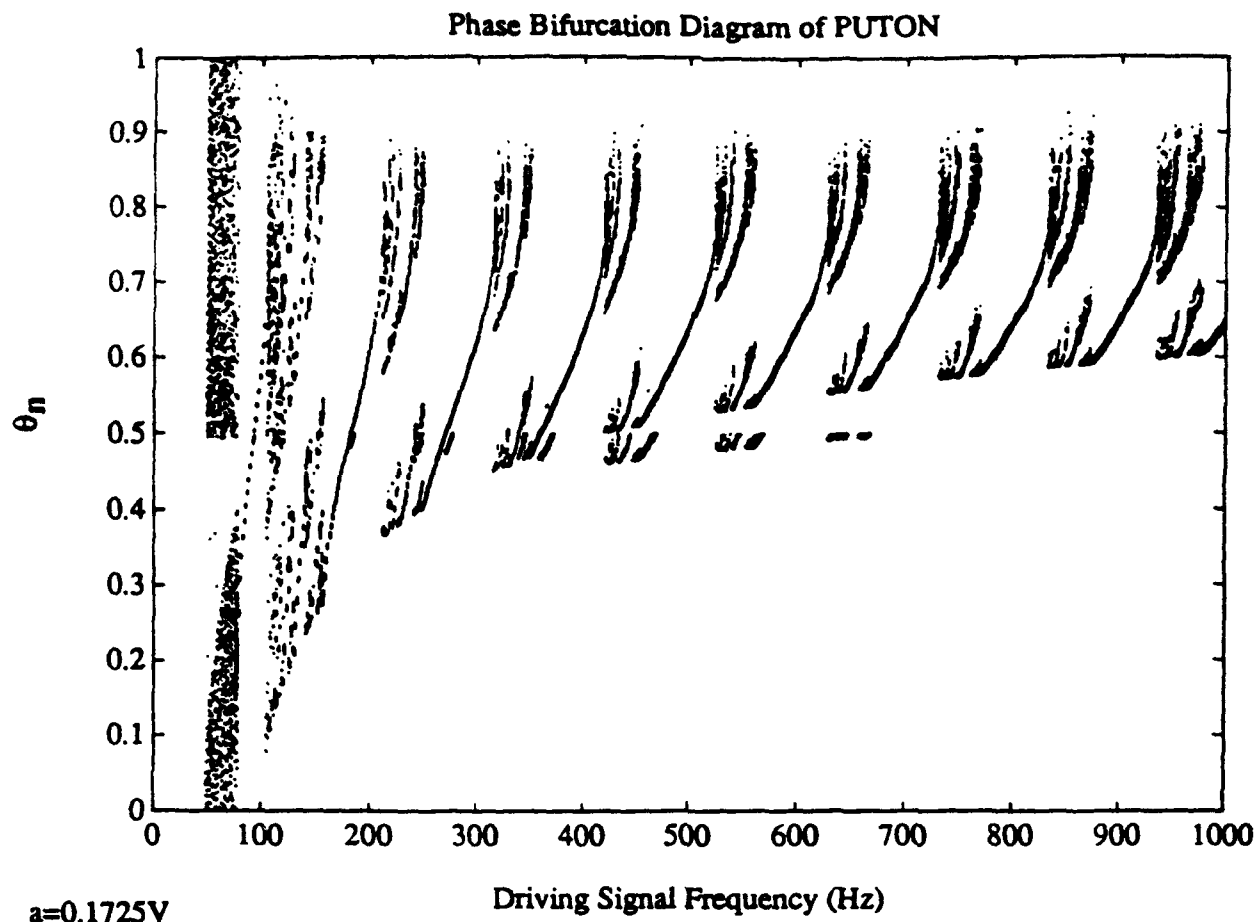


Fig. 10. Measured bifurcation diagram for a PUTON embodiment of the bifurcating neuron for  $a = .175$  [V]. Note the variety of phase-locked regular firing modalities for which only one value of  $\theta_n$  is produced at each driving frequency, separated by more complicated and perhaps erratic  $\theta_n$  sequences occurring at other driving frequency values.

The information in the bifurcation diagram can also be presented in a somewhat different format, that of a *Sawtooth Phase-lock diagram* or its equivalent the *Devil's Staircase diagram*. Since  $\theta_n$  is the relative phase between the  $n$ -th spike fired by the neuron and the immediately preceding peak of its cosinusoidal activation potential or driving signal, the timing of the spikes can be determined and therefore also the interspike interval or its inverse: the instantaneous firing frequency  $f$ . Plots of the firing frequency  $f$  vs. driving frequency  $f_s$  or of the *phase-lock ratio*  $\rho = \frac{f_s}{f}$  vs.  $f_s$  also

called the *rotation ratio* computed from the  $\theta_n$  sequences are shown in Fig. 11 for a bifurcating neuron assuming a glow-lamp nonlinearity. Note both phase lock diagrams show presence of driving signal frequency windows over which perfect phase-locking with negligible standard deviation in firing frequency occur. These regions are separated by regimes of more complex firing modalities including quasi-periodic perhaps erratic firing for which the standard deviation of the firing frequency is finite. The Devil's Staircase diagram measured for a bifurcating neuron embodiment employing a PUT nonlinearity is shown in Fig. 12(a) with a high resolution (expanded) plot of one of the segments lying between two phase-lock regions given in Fig. 12(b). This latter plot shows clearly the self-similar or fractal nature of the diagram where a staircase structure is easily discerned in the mean and the mean plus and mean minus standard deviation branches of the diagram. Note however that the finer rungs of the staircase appear to be blurred by noise in the PUTON circuit used in making the measurement. The transition from a phase-lock region where the standard deviation of firing frequency (or rotation number) is negligible, to a region where the standard deviation is finite is seen to be quite abrupt attesting to the rapid switching of behavior of the bifurcating neuron as the value of the bifurcation parameter,  $f_s$  in this case, is altered. Again the plots of Figs. 11 and 12 are for a fixed value of driving signal amplitude. These plots can be regarded as the *phase-lock frequency response* of the bifurcating neuron at fixed driving signal amplitude. Phase-lock frequency response plots at different discrete values of driving signal amplitude can be obtained in a similar fashion. The data contained in such a series of plots can be presented compactly in the form of the *Arnold Tongues' diagram* shown in Fig. 13. This diagram can be interpreted as the frequency response of an active nonlinear device, like the bifurcating neuron, that is capable of phase-locking its firing to the driving signal. The horizontal lines in this plot give the width of the periodic phase-lock firing regions at each driving signal amplitude as the frequency is swept. Outside these regions, which are shaped like wedges or tongues pointed downward, the firing can be period- $N$ , quasi-periodic or even erratic. The tongues represent phase-lock regions with, starting from the left, integer phase-lock ratio  $\rho = 1, 2, 3, \dots$ . It is noted the width of these integer phase-lock regions decreases as the amplitude of the driving signal is made smaller.

In deriving the PTM, we assumed cosinusoidal periodic activation or driving signal of the neuron to make the analysis tractable. We believe the complex behavior of the bifurcating model neuron observed for cosinusoidal activation does persist for arbitrary periodic activation. Replacing the cosinusoidal driving signal in Fig. 6 by an arbitrary periodic signal hints that this should be so and preliminary simulation results for an I & F neuron driven by a periodic signal composed of several sinusoidal components shows interspike sequences with similar complexity as when it is driven by a pure cosinusoidal signal. An interesting question in this regard is

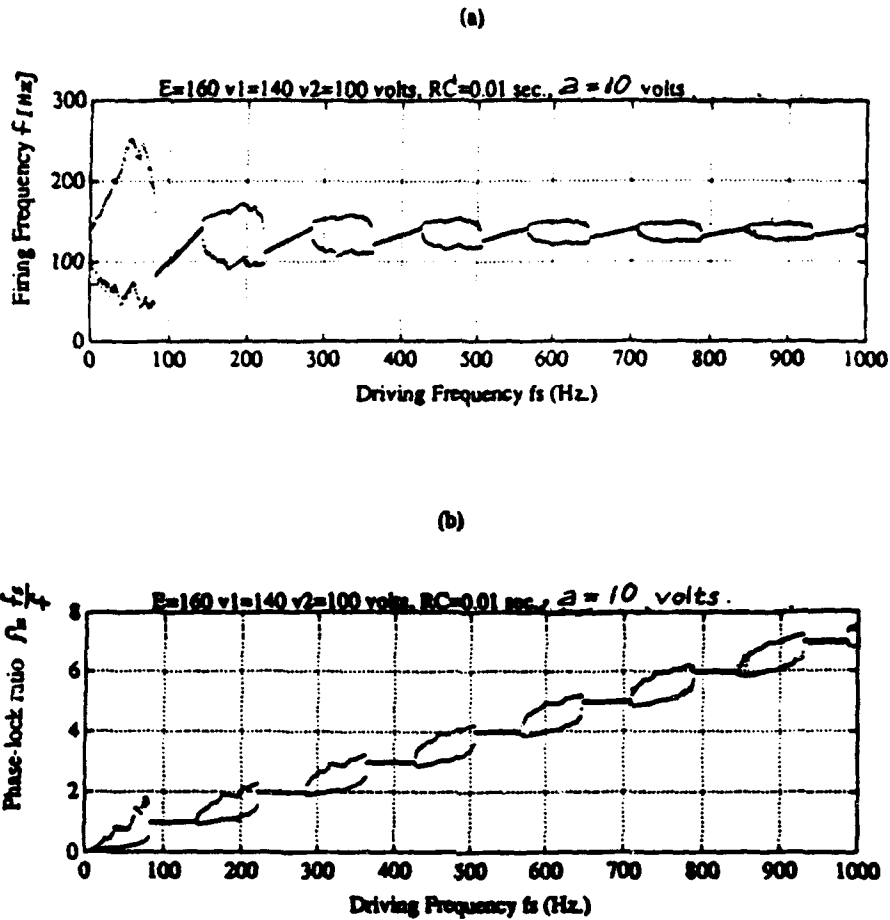


Fig. 11

Computed *phase-lock frequency response* of bifurcating neuron employing glow-lamp S-shaped nonlinearity. Driving cosinusoidal signal is  $u(t) = a \cos(2\pi f_s t)$ . Plots shown are for  $a = 10[V]$  and operating conditions: supply voltage  $E = 160 [V]$ , glow-lamp breakdown and extinction voltages  $V_1 = 140[V]$ , and  $V_2 = 100[V]$  respectively. a) Sawtooth phase-lock diagram and b) corresponding staircase phase-lock diagram. In both diagrams the upper branches are the mean plus standard deviation of the firing frequency and the lower branches are the mean minus the standard deviation of the firing frequency. In the phase-lock regions the standard deviation is zero. The mean and standard deviation at each driving frequency  $f_s$  are calculated from a 200 msec record of the firing activity. Note in both plots the integer phase-lock regions (where the phase-lock ratio  $\rho = f_s/f = 1, 2, 3, \dots$ ) are separated by regions of irregular firing where the standard deviation in firing frequency is finite. These regions of erratic firing furnish adaptive "noise" that could play a role in "annealing" bifurcating neuron networks i.e., drive them into entrained (phase-locked) states.

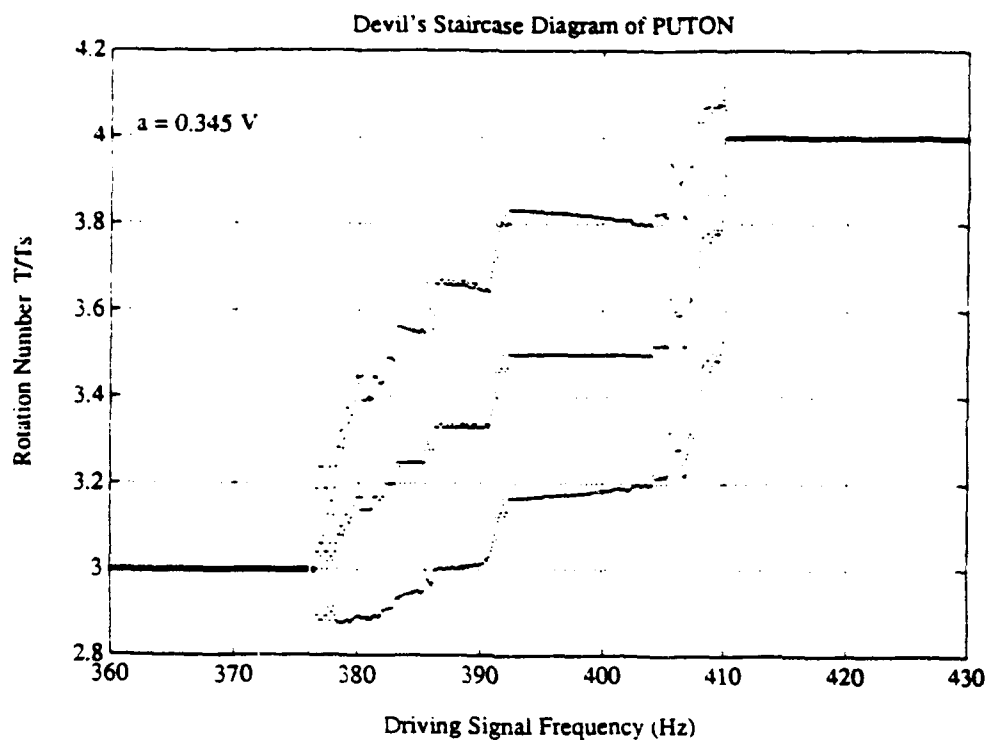
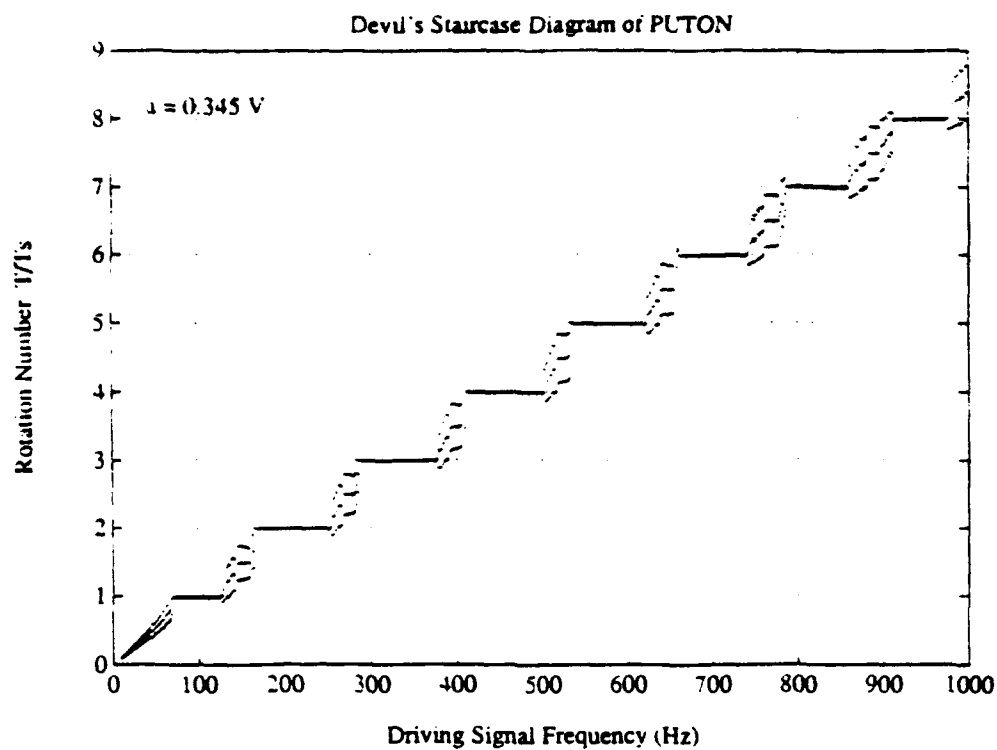


Fig. 12. Measured Devil's Staircase diagram of the periodically driven PUTON, a bifurcating neuron embodiment employing a programmable unijunction transistor (PUT) furnishing the S-shaped membrane nonlinearity (a), and expanded region of the diagram illustrating the self-similar or fractal nature of the Devil's Staircase diagram (b).



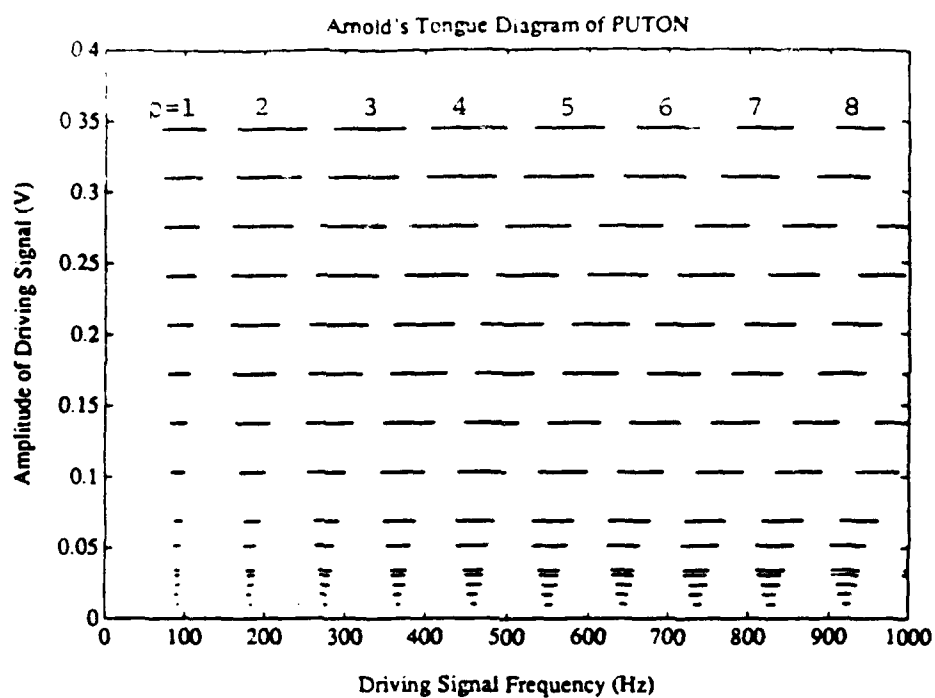


Fig. 13. The Arnold Tongue's diagram of the PUTON representing the phase-lock regions of the periodically driven PUTON (horizontal bars) for selected driving signal amplitude. If standard deviation of the interspike interval is below .04 [ms] a horizontal bar is plotted. In the regions between the horizontal bars, which form the tongues, the firing is quasi-periodic or erratic.

whether the bifurcating neuron phase-locks its firing to the strongest spectral component of the arbitrary periodic signal or not. This question is under investigation.

### **3. IMPLICATIONS AND SIGNIFICANCE OF THE BIFURCATING NEURON CONCEPT.**

The results presented so far show that the model neuron we consider has much more complex behavior than simple sigmoidal response employed widely in neural modeling today, where the state of the neuron is represented by its firing frequency, and the firing frequency is a nonlinear (sigmoidal shaped) function of the activation potential. In this mode of operation, and when the activation potential is above threshold, the neuron can be viewed as a VCO (voltage controlled oscillator) with a highly nonlinear (spike) output waveform. We have seen, however, when the activation potential becomes periodic, the neuron can alter its behavior (bifurcate) to fire in a variety of phase-locked regular or erratic firing modalities depending on the parameters (amplitude and frequency) of the activation potential. Hence our use of the term *bifurcating neuron*. Thus the bifurcating neuron can act not only as a sigmoidal neuron, as when its activation potential is slowly varying and is not periodic, but also can act as a detector/encoder of episodes of coherent incident spike wavefronts at its dendritic-tree that give rise to periodic activation.

The PTM, which is derived in the form given here for the first time, provides a powerful tool for studying the dynamics of realistic neurons and neural networks using the customary tools of nonlinear dynamics such as: bifurcation diagram, Lyapanov exponent, entropy of the firing patterns or phase orbits, Devil's staircase diagram, and Arnold Tongues' diagram, all tools that provide a novel and useful approach to modeling, characterizing, and better understanding of cortical networks, higher-level cortical functions, and of how to ultimately incorporate feature-binding and cognition and other higher-level cortical functions in man-made systems.

In the above analysis a clear relation between a spike or action potential fired by a neuron and the limit-cycle trajectory in the  $i$ - $v$  plane of the S-shaped nonlinearity used to represent the neuron's excitable membrane was established. It is interesting to speculate as to why use of limit-cycle oscillations and spiking neurons has evolved in biological systems. We offer the following reasons for such use:

- (a) Stability of the limit-cycle oscillation, and hence its robustness and immunity while it persists to noise: The shape of spikes fired by the neuron is invariant.
- (b) Rapid entrainment and synchronization of systems possessing limit-cycle oscillations such as the bifurcating neuron which would allow coherent states to evolve rapidly within a coupled population of bifurcating neurons (bifurcating neural network). A coherent or synchronized state is manifested by fixed relative-phase pattern (vector) between the firings of neurons in the network and by coherent incident spike wavefronts received by each neuron from other neurons in the network.
- (c) The relative-phase vector of a synchronized network of  $N$  neurons is defined by an  $N$ -vector  $\vec{\phi}$  with elements  $\phi_i$   $i=1,2,...,N$  where  $0 \leq \phi_i < 2\pi$  is the phase between a spike

fired by the  $i$ -th neuron and the immediately preceding spike of an arbitrarily selected reference neuron. If the  $i$ -th neuron fires  $\tau_i$  [sec] after the reference neuron, then  $\phi_i = \frac{\tau_i}{T} 2\pi$  where  $T$  is the interspike interval of the reference neuron. By this definition, the reference neuron's in phase would be zero. If, because of measuring accuracy, the phase of each neuron can be distinguished over  $L$  levels then the total number of possible distinguishable permutations or states the network can assume is  $(NL)!$ . In contrast, a network of  $N$  binary neurons has a total of  $2^N$  possible states while a network of  $N$  sigmoidal neurons whose outputs are distinguishable over  $L$  levels has a total of  $L^N$  possible states. If we take for the sake of illustration,  $N=3$  and  $L=3$ , then the bifurcating neuron network can have a total of  $(3 \times 3)! = 362,880$  possible states. This exceeds by far the  $2^3=8$  states possible had the network been of binary neurons or the  $3^3=27$  states had it used sigmoidal neurons. These simple considerations serve to illustrate that in their making use of phase information, temporal networks, possess far denser state-space available for accessing by network dynamics than conventional sigmoidal or binary neuron networks. Accordingly one can intuitively expect that bifurcating neural networks would exhibit richer state-space behavior than sigmoidal networks or their binary cousins and to be capable, in general, of carrying out more complex signal processing operations and computations.

It is interesting to note, in connection with the above remarks that a possible advantage of multilevel sigmoidal neuron networks demonstrated in handling a gray-level image processing application is that the number of neurons and number of interconnections are reduced compared to binary neuron networks [21] which is an important concern in VLSI implementations of neural network. In the referenced work each neuron's response was represented by a multilevel sigmoidal function of  $L=16$  levels. The results suggest that the smaller the slope of the staircase nonlinearity for each neuron, i.e., the larger is the value of  $L$ , the larger is the domain of attraction of each desired equilibrium point. The complexity of VLSI realization of multilevel threshold elements can neutralize however the advantage of reduced number of neurons and interconnections in such networks. Since the number of possible state levels (relative phase values) that can be naturally and automatically assumed by a neuron in a synchronized bifurcating or spiking neural network can be quite large, such networks could possess a distinct advantage, over multilevel threshold neural networks, in reducing the number of neurons and interconnections needed to handle information processing tasks. This is an interesting subject for further study of the practical advantages of bifurcating neural networks.

- (d) If we adopt the relative-phase vector, as the state variable for temporal networks, it would be tempting to speculate that relative-phase patterns in cortical networks serve as substrates for cognition and other higher-level cortical functions. A relative-phase vector can be represented as a point in a "relative-phase" state-space of the population or network. Periodic changes in the relative-phase pattern or vector in such networks can then be represented by a closed trajectory or limit-cycle in relative-phase state-space of the network. Similarly a chaotic or erratic sequence of relative-phase vectors can be associated with a chaotic trajectory in the relative-phase state-space of the network. The relative-phase state-space of bifurcating neural networks could thus exhibit point, periodic, and chaotic

attractors. In this picture, chaos and the possibility of rapid bifurcation between such attractors, as induced by external signals, is of interest as means for rapidly searching the state-space of the network for coherent states that are meaningful for cognition, feature-binding, and other kinds of higher-level processing operations believed to be carried out by the cortex.

Defining the relative-phase pattern of a synchronized network of bifurcating or spiking neurons in the above manner is ambiguous because of the arbitrariness of choosing the reference neuron. To remove this ambiguity we recall that in the analysis of Section 2, the relative-phase of the spikes fired by the individual neuron in a synchronized network was defined relative to the peak, or some other feature, e.g., zero-crossing, of the periodic activation signal driving it. Thus the periodic activation signals produced at the hillock of each neuron in a synchronized network can furnish a natural self-reference for determining the relative phase of spikes produced by each neuron and hence the relative-phase distribution of the network at any time without ambiguity. Of course, the self-reference signals exist only when the network is in synchronized state. Accordingly the relative phase vector is  $\bar{\theta}(t)$  whose components  $0 \leq \theta_i(t) < 2\pi$ ,  $i=1,2,\dots,N$  with  $N$  being the number of neurons of the network, are the relative-phase vs. iteration number, i.e., the orbits, produced by the neurons. The relative-phase vector  $\bar{\theta}(t)$  represents now the state of the network unambiguously. At any instant of time,  $\bar{\theta}(t)$  describes the position vector of a point in an  $N$ -dimensional relative-phase state-space of the network whose coordinates span the  $[0, 2\pi]$  range.

Thus when the orbit or sequence  $\theta_n$   $n=1,2,3,\dots$  for each neuron is fixed i.e.,  $\theta_n = \text{const}$  for each neuron, the relative-phase pattern of the synchronized network will be fixed i.e.,  $\bar{\theta}(t)$  is constant and the behavior of the network is represented by a fixed point in state-space. In this case the network is both phase-locked and frequency locked. When  $\bar{\theta}(t) = \bar{\theta}(t + T)$  i.e., the state vector evolution in time is periodic, the behavior is represented by closed state-space trajectory. The network assumes the same value of the relative-phase vector every  $T$  seconds. In this condition, every neuron can be in a period- $m$  firing modality with the ratios of the values of  $m$  for the various neurons being related by integer numbers. The network are then phase-locked but not frequency locked.

A network of bifurcating neurons in which the neurons can exhibit quasi-periodic or erratic firing for certain parameters of their periodic activation signals, could exhibit more complicated state-space trajectories. Such a network would contain neurons with erratic firing whose number and identity can change in time producing therefore quasi-periodic and chaotic state-space trajectories that could visit large regions of the state space. This suggests that such chaotic states or trajectories could serve as means for searching the relative-phase state-space of the network for point or periodic attractors that could represent meaningful cognitive states or could, as proposed by Zak [22], represent higher-level cognitive processes such as formation of new logical forms based upon generalization and abstraction.

- (e) **Dynamic Partitioning:** There is mounting evidence at present that the signal processing function of the dendritic-tree of a living neuron is not confined to merely integrating the post synaptic potentials produced by its synaptic inputs (incident spike wavefront) in order to form the activation potential at the neuron's hillock, but could also include more complex nonlinear signal processing operations [23,24]. These operations are believed to stem from the action of excitable membranes at some spines on the dendritic-tree which makes for an active rather than passive dendritic-tree. This could give rise, in a coupled population of such dendritic-neurons, to stimulus driven synchronization and feature linking capabilities [25], and stimulus driven dynamic partitioning of a network into weakly interacting subpopulations [15],[26]-[28] that can perform collective computations in parallel which is significant for forming Non Lipschitzian networks with unpredictable dynamics. It is suggested [2] that Non Lipschitzian networks represent cortical networks better than conventional neural networks whose dynamics obey the Lipschitz condition. Dendritic-tree processing is meaningful in spiking neuron networks and therefore it is not an issue in sigmoidal neuron networks. Therefore consideration of spatiotemporal processing operations in dendritic-trees does not arise in sigmoidal networks.
- (f) The spiking nature of the bifurcating neuron and the complexity of phase orbits ( $\theta_n$  values associated with the spike trains) it produces under changing input conditions enables viewing the bifurcating neuron as an *information source*. It also enables defining the *entropy* and *mutual information* of the  $\theta_n$  sequences or firing patterns produced. Specifically, we can view the bifurcating neuron as an information source with sequential output of symbols or events from the set  $S = \{S_1, S_2, \dots, S_q\}$  with each symbol occurring with fixed probability  $P(S_i)$   $i=1,2,\dots,q$ . If the probability of a symbol occurring is independent of previous symbols we say the neuron is a zero-memory source. The information gained or received when the  $i$ -th symbol or event occurs is then by definition [29],

$$(29) \quad I(S_i) = \log_2 \frac{1}{P(S_i)}$$

The average amount of information received per symbol is

$$(30) \quad \langle I \rangle = \sum_{i=1}^q P(S_i) I(S_i)$$

$$(31) \quad = - \sum_{i=1}^q P(S_i) \log_2 P(S_i) = H$$

where  $H$  is the entropy.

We need to be more specific about what is meant by symbol when we view the bifurcating neuron as an information source. Consider the plot  $\theta_n$  vs.  $n$  shown in Fig. 14 where  $n=1,2,\dots,N$ .

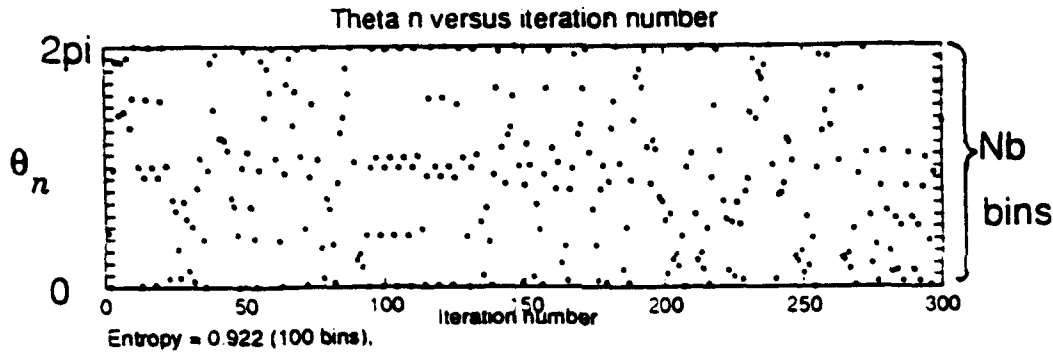


Fig. 14. Plot of typical  $\theta_n$  vs. iteration number  $n$  referred to in entropy calculations.

There are  $N$  points in the plot and we assume  $N$  is sufficiently large. The  $\theta_n$  axis, spanning the  $0$ - $2\pi$  interval, is divided into  $N_b$  bins. Then, by counting how many points fall in the  $i$ -th bin and dividing the outcome by  $N$ , we obtain the probability  $P_i$  of  $\theta_n$  falling in bin  $i$ . This is also the frequency histogram of  $\theta_n$ .

The entropy, which is a measure of disorder, is in accordance to eq. (31),

$$H \equiv - \sum_{i=1}^{N_b} P_i \log P_i$$

The maximum entropy or disorder occurs when the probability of  $\theta_n$  falling in the  $i$ -th bin ( $i=1,2,3,\dots,N_b$ ) is constant. The number of events per bin in this case is  $\frac{N}{N_b}$  and  $P_i = \left(\frac{N}{N_b}\right) \times \left(\frac{1}{N}\right) = \frac{1}{N_b}$ . Thus

$$H_{\max} = - \sum_{i=1}^{N_b} P_i \log (P_i) = - \sum_{i=1}^{N_b} \left(\frac{1}{N_b}\right) \log \frac{1}{N_b} = \log N_b$$

Therefore the normalized Entropy

$$(32) \quad H_n = \frac{H}{H_{\max}} = - \frac{1}{\log N_b} \sum_{i=1}^{N_b} P_i \log P_i$$

provides a normalized measure of the degree of disorder in the orbits  $\theta_n$  and therefore in spike trains produced by the bifurcating neuron. An  $H_n = 1$  indicates chaotic firing. The computed normalized entropy of the  $\theta_n$  sequence in Fig. 14 for example, assuming  $N_b = 100$  bins is  $H_n = .922$ .

We have seen in the above that the bifurcating neuron can be viewed as an information source producing symbols  $\theta_n$   $n=1,2,\dots$  with symbol probability  $P(\theta_n)$ . Because the orbit  $\theta_n$  depends on the neuron's input (activation potential), the symbol probability changes as the input to the neuron, or the incident spike wavefront giving rise to it, changes. Different inputs produce therefore different symbols and corresponding symbol probabilities. The bifurcating neuron can therefore also be viewed as a complex *encoder* of incident spike wavefronts.

The above observations may enable speculating on possible learning in bifurcating neural network in terms of reducing the *Cross-Entropy*,

$$(33) \quad G = \sum_{n=1}^N P(\theta_n) \log \frac{P'(\theta_n)}{P(\theta_n)}$$

where  $P(\theta_n)$  is the symbol probability produced by the neuron for certain input (incident spike wavefront) and  $P'(\theta_n)$  is a target symbol probability we wish the neuron to produce for that input. The learning task then is how to modify the synaptic responses (synaptic weights and possibly time constants or synaptic delays) of the neuron so that  $G$  is minimized. Note that  $G \rightarrow 0$  when  $P'(\theta_n) \rightarrow P(\theta_n)$ . The goal then is to determine the synaptic responses of neurons in a population or pool of interconnected bifurcating neurons in which a portion of the neurons receive external inputs such that given input patterns which can be spatial or spatio-temporal, end up producing

desired distinct relative phase patterns when the network gets synchronized. The learning algorithm developed must obey the usual minimal perturbation principle, i.e., reoccurrence of input patterns that have already been learned would not alter the synaptic weights and only novel inputs would produce synaptic response changes that minimally perturb the information already stored.

(g) Although in preceding sections we refer to erratic firing of our bifurcating model neurons under certain conditions of periodic activation, the Lyapunov exponents associated with values of  $a$  and  $f_s$  in the Arnold Tongues' diagram falling between the phase-lock regions have never been found to be positive which precludes explicit chaotic firing. The seemingly erratic firing regions in the bifurcation diagram would therefore be representing period- $m$  firing with long periods, i.e., large  $m$  or equivalently firing modalities where a long sequence of non-repeating values of  $\theta_n$  occurs before the value of  $\theta_n$  is repeated. This observation is supported by the Poincaré-Bendixon theorem [30] which says that a system whose nonlinear dynamics are governed by two autonomous first-order ordinary differential equations containing nonlinear coupling terms cannot exhibit chaotic behavior (see also pages 2-5 of reference 20). The behavior of the cosinusoidally driven bifurcating neuron model analyzed in Section 2 is governed by two differential equations:

$$(34) \quad \frac{dv}{dt} = \frac{E - v}{R} - \phi(v - u)$$

and

$$(35) \quad \frac{d\phi}{dt} = \omega$$

where  $u = a \cos\phi$  and  $\phi = \omega t + \theta_0$ ,  $\theta_0$  being a constant and action potentials are produced whenever  $v$  reaches threshold (see eq. 5). According to the Poincaré Bendixon theorem this system of equations cannot exhibit chaotic solutions i.e., chaotic sequences of firing phases  $\theta_n$ . It is capable of producing as we stated long sequences of period- $N$  that appear erratic to an observer if  $N$  is very large but are not strictly chaotic. Therefore in order to exhibit chaos, our bifurcating neuron model must be modified in such a way as to make its dynamics governed by at least three autonomous first-order differential equations with nonlinear coupling terms. This could occur when, for example, the threshold level  $v_1$  of the S-shaped nonlinearity  $i = \phi(v)$  is not constant, as assumed in the analysis of Section 2, but obeys a differential equation of its own like,

$$(36) \quad \frac{dv_1}{dt} = -\alpha v_1 + f(v)$$

where  $\alpha$  is a constant and  $f(v)$  is a nonlinear function of  $v$ . Equation (36) adds *adaptive thresholding* or *accommodation* to our bifurcating neuron model making it resemble more



biological neurons which are known to exhibit accommodation. It is worth noting that the Hodgkin-Huxley model of the biological membrane [18] and the MacGregor and Oliver version of it [31] consist of  $N \geq 3$  autonomous first-order differential equations with nonlinear coupling terms. They are expected therefore to exhibit chaotic solutions for certain regions of their parameter spaces. It would seem therefore that under suitable conditions of periodic activation, biological neurons in the cortex would exhibit chaotic firing. Indeed chaotic firing regimes have been observed experimentally in the periodically driven biological membrane (the squid's giant axon) and the Molluscan neuron [33]-[36].

Inclusion of adaptive thresholding or accommodation in the analysis of the bifurcating model neuron described in this paper, although analytically challenging, is very well likely to lead to complex firing modalities under periodic activation that include full blown chaotic firing. One may well ask at this point why is chaotic firing important? The following is an attempt to answer this question.

Chaos describes a strange intermediate state lying between rigid organization, i.e., order, and complete disorganization, i.e., disorder or randomness. It also connotes something between predictability and chance, between a deterministic signal and noise.

Chaos in neurodynamics could have beneficial consequences because decision making processes that show chaos mix consistency with unpredictability and could overcome the limitations of both.

Understanding chaos and learning to use it in the design of in man-made systems in general and in neurodynamics in particular could be the key for increasing the power of neurocomputing by extending their capabilities to higher-level processing like cognition, feature and concept binding, inferencing, reasoning, attention focusing, and possibly improved learning and optimization.

Possible roles for chaos are:

- Efficient search of state-space of a network that is neither systematic or random, but is driven by the dynamics of the network itself.
- As inherent (self-generated) adaptive (signal dependent) source of "noise" in a network, chaos could possibly furnish a mechanism for "self-annealing" the network into phase-locked states of "energy" minima that are cognitively meaningful.

**4. CONCLUSIONS.** The study of bifurcating and spiking neural networks and the roles of synchronicity, bifurcation and chaos in such networks is still in its infancy. It is however rapidly growing because of its promises to provide better understanding of higher-level cortical functions and of how to incorporate them in artificial neural networks to enhance the power of neurocomputing. Recent developments [25],[37] are examples of this trend. The study of the dynamics of the bifurcating model neuron presented here is a step towards understanding how the functional complexity of the individual processing element governs the dynamics of bifurcating

neural networks specially when they enter synchronized states and towards learning how to use synchronicity bifurcation and chaos in the design of a new generation of computing structures.

**5. ACKNOWLEDGEMENT.** The research described was sponsored by the Army Research Office and SDIO/IST under ONR management.

## **6. REFERENCES.**

1. N. Farhat and M. Eldefrawy, "The Bifurcating Neuron: Characterization and Dynamics", Proc., *SPIE*, vol. 1773, pp. 23-34, 1993.
2. M. Zak, "An Unpredictable-Dynamics Approach to Neural Intelligence", *IEEE Expert*, pp. 4-10, August 1991.
3. Chr. v.d. Marlsberg, Internal Report 81-2, MPI Biophys. Chem. Gotingen, 1981.
4. M. Abeles, Local Cortical Circuits, Springer Verlag, Berlin, 1982.
5. C.M. Gray and W. Singer, *Soc. Neuroscience Abstract* 12404.3, 1987.
6. C.M. Gray, et. al., *Europ. J. Neurosc.*, vol. 2, pp. 607-619, 1990.
7. A.K. Engel, et. al., *Europ. J. Neurosc.*, vol. 2, pp. 588-606, 1990.
8. C.M. Gray, et. al., *Nature*, vol. 338, pp. 334-337, 1989.
9. R. Eckhorn, et. al., *Biol. Cybern.*, vol. 60, pp. 121-130, 1988.
10. P. Konig and T. Schillen, *Neural Computation*, vol. 3, pp. 133-166, 1991.
11. \_\_\_\_\_, *Neural Computation*, vol. 3, pp. 167-178, 1991.
12. W. Freeman, "The Physiology of Perception," *Sci. Am.*, pp. 78-85, Feb., 1991.
13. S. Blackeslee, "Nerve Cell Rythm may be Key to Consciousness," *The New York Times* (Science Times Section), Oct. 27, 1992.
14. K. Delaney, et. al., "Waves and Stimulus-Modulated Dynamics in an Oscillating Olfactory Network", (Private comm.)
15. E.R. Grannan, D. Kleinfeld, and H. Sompolinsky, "Stimulus-Dependent Synchronization of Neural Assemblies", *Neural Computation* vol. 5, pp. 550-569, July 1993.
16. L. Shastri and V. Ajjanagadde, "From Simple Associations to Systematic Reasoning", *Behavioral and Brain Sciences*, (To be published, September, 1993).

17. P.M. Koch, "Microwave Ionization of Excited Hydrogen Atoms: What We Do Not Understand", in Chaos, D.K. Campbell (Ed.), AIP, New York, p. 445, 1990.
18. A.L. Hodgkin and A.F. Huxley, "The Components of the Membrane Conductances in the Giant Axon of *Loligo*", *J. Physiol.*, vol. 116, pp. 473-496, 1952.
19. A.A. Andronow and C.E. Chaikin, Theory of Oscillations, Princeton University Press, Princeton, NJ, p. 171, 1949.
20. G.L. Baker and J.P. Gollub, Chaotic Dynamics, Cambridge University Press, Cambridge, pp. 85-89, 1990.
21. J. Si and A. Michel, "Analysis and Synthesis of Discrete-Time Neural Networks with Multilevel Threshold Functions", Proc. 1991 IEEE Int. Symp. on Circuits and Systems, vol. 3 (of 5) on Analog Circuits and Neural Networks, IEEE Publication # 91 CH3006-4, 1991.
22. M. Zak, "Chaos as a Part of Logical Structure in Neurodynamics", *App. Math. Lett.*, vol. 2, pp. 175-177, 1989.
23. W. Rall and I. Segev, "Excitable Dendritic Spine Clusters: Nonlinear Synaptic Processing", in Computer Simulation in Brain Science, R.M.J. Cotterill (Ed.), Cambridge Univ. Press, Cambridge, pp. 26-43, 1988.
24. John G. Elias, "Artificial Dendritic-Trees", *Neural Computation*, vol. 5, pp. 648-664, July, 1993.
25. R. Eckhorn, H.J. Reitboek, M. Arndt, and P. Dike, "Feature Linking via Synchronization among Distributed Assemblies: Simulations of Results from the Cat Visual Cortex", *Neural Computation*, vol. 2, pp. 293-307, Fall, 1990.
26. A. Nischwitz, H. Glunder, A. van Oertzen and P. Klausner, "Synchronization and Label-Switching in Networks of Laterally Coupled Neurons", Artificial Neural Networks, 2, I. Aleksander and J. Taylor (Eds.), Elsevier Science Publishers, B.V., pp. 851-854, 1991.
27. I. Segev, "Computer Study of Presynaptic Inhibition Controlling the Spread of Action Potentials into Axonal Terminals", *J. of Neurophysiology*, vol. 63, pp. 987-998, May, 1990.
28. M. Usher, H.G. Schuster, E. Niebur, "Dynamics of Populations of Integrate-and-Fire Neurons, Partial Synchronization, and Memory", *Neural Computation*, vol. 5, pp. 570-586, July, 1993.
29. J.A. Freeman and D.M. Skapura, Neural Networks, Addison-Wesley, Reading, MA, (Section 5.1), 1991. See also, P.M. Woodward, Probability and Information Theory with Applications to Radar, (Second edition), Pergamon Press, Oxford, (Chapter 3), 1953.

30. F.C. Hoppensteadt, Analysis and Simulation of Chaotic Systems, Springer-Verlag, New York, pp. 31-32, 1992.
31. R. MacGregor and R. Oliver, "A Model for Repetitive Firing in Neurons", *Kybernetik*, vol. 16, pp. 53-64, 1974.
32. R. Guttman, L. Feldman and E. Jacobsson, "Frequency Entrainment of Squid Axon Membrane", *J. Membrane Biology*, vol. 56, pp. 9-18, 1980.
33. A.V. Holden and S.M. Ramadan, "The Response of a Molluscan Neurone to Cyclic Input: Entrainment and Phase-Locking", *Biol. Cybern.*, vol. 43, pp. 157-163, 1981.
34. A.V. Holden, et. al., "The Induction of Periodic and Chaotic Activity in a Molluscan Neurone", *Biol. Cybern.*, vol. 43, pp. 169-173, 1982.
35. G. Matsumoto, K. Aihara, M. Ichikawa, and A. Tasaki, "Periodic and Nonperiodic Responses of Membrane Potential in Squid Giant Axon Under Sinusoidal Current Stimulation", *J. Theor. Neurobiol.*, vol. 3, pp. 1-14, 1983.
36. K. Aihara and G. Matsumoto, "Chaotic Oscillations and Bifurcation in Squid Giant Axon", in Chaos, A.V. Holden (Ed.), Princeton Univ. Press, Princeton, NJ, pp. 257-269, 1986.
37. J.L. Johnson, "Pulse-Coupled Neural Nets: Translation, Rotation, Scale, and Intensity Signal Invariant for Images", (Submitted to *App. Optics*, July 13, 1993).

# SHOCK INDUCED SURFACE INSTABILITIES AND NONLINEAR WAVE INTERACTIONS

BRIAN BOSTON, JOHN W. GROVE<sup>1,2,3,5</sup>, RICHARD HOLMES, L. F. HENDERSON<sup>1</sup>,  
DAVID H. SHARP<sup>7</sup>, YUMIN YANG<sup>4</sup>, AND QIANG ZHANG<sup>4,6</sup>

**ABSTRACT.** We discuss the application of front tracking to the simulation of shock reflections and shock accelerated interfaces. Some key features of the front tracking method are the elimination of numerical diffusion and the reduction of wall heating. In computations of the regular Mach reflection of a shock at an oblique ramp, we see enhanced resolution of the primary waves in the interaction. In addition, tracking allows very precise measurements to be made of the states and location of the Mach triple point. Our computations of the growth rate of a Richtmyer-Meshkov unstable interface are the first numerical results that are in quantitative agreement with experimental results of a shocked air-SF<sub>6</sub> interface. Previous attempts to model the growth rate of the instability have produced values that are almost twice that of the experimental measurements. Moreover, the failure of the impulsive model, and the linear theory from which it is derived, to model experiments correctly is understood in terms of time limits on the validity of the linear model.

**Keywords:** Front tracking, Mach reflection, Richtmyer-Meshkov

## 1. INTRODUCTION

In this article we present results of simulations using front tracking combined with a second order Godunov finite difference method. Two classes of problems are discussed, the oblique reflection of shock waves at ramps, and the computation of the instability growth rate of a perturbed, shock-accelerated interface. Our code achieves excellent resolution of the simulated flows, even on the relatively coarse grids used here. In both cases our computed results are shown to be in excellent agreement with experiments. Indeed, we present computations of the Richtmyer-Meshkov instability that for the first time agree with experimentally measured growth rates of interface perturbations.

The ramp reflection simulations model the interaction of a planar shock wave with an oblique wall. We are interested in determining the structure of the reflection process for ramp angles that are very close to the mechanical equilibrium condition for bifurcation to regular reflection, as defined by the coincidence of regular and Mach reflection. The use of front tracking allows us to conduct numerical experiments that are extremely close to this point.

---

<sup>1</sup>Supported in part by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University under subcontract to SUNY Stony Brook, ARO contract number DAAL03-91-C-0027.

<sup>2</sup>Supported in part by the National Science Foundation Grant no. DMS-9201581

<sup>3</sup>Supported in part by the U. S. Army Research Office, grant no. DAAL03-92-G-0185.

<sup>4</sup>Supported in part by the Applied Mathematics Subprogram of the U.S. Department of Energy DE-FG02-90ER25084.

<sup>5</sup>Supported in part by the National Science Foundation Grant no. DMS-9057429

<sup>6</sup>Supported in part by the Oak Ridge National Laboratory subcontract 19X-SJ067V.

<sup>7</sup>Supported by U. S. Department of Energy

Ordinary shock capturing methods are unable to resolve the Mach triple point configuration in this regime, due to the extreme closeness of this point to the wall. We measure several quantities for the reflection process, including the trajectory of the triple point, and the Mach number of the flow behind the foot of the Mach stem. Comparison of these quantities to experiment shows a good agreement between our values and the experimentally measured quantities.

The Richtmyer-Meshkov instability concerns the growth of interface perturbations on a shock accelerated material interface. When a shock wave collides with the interface between two different materials, small perturbations of this interface grow into nonlinear structures having the form of "bubbles" and "spikes". The occurrence of this shock-induced instability was predicted by Richtmyer [15] and confirmed experimentally by Meshkov [12]. The Richtmyer-Meshkov instability is similar to the more familiar Rayleigh-Taylor instability and is important in both natural phenomena (supernovae) and technological applications (inertial confinement fusion).

Theory and computation have so far failed to provide an understanding of the Richtmyer-Meshkov instability that is in quantitative agreement with existing experiments [3, 4, 6, 14, 16]. Computations of the Richtmyer-Meshkov instability for singly shocked, sinusoidally perturbed interfaces have over-predicted growth rates by factors from 40% to 100% [6] as compared to experiments. The main theoretical model used in this area, Richtmyer's impulsive model [15], also consistently predicts a growth rate that is too large.

Our computations of the Richtmyer-Meshkov instability are further validated by a comparison of small amplitude perturbation, early time simulations with solutions to a linearization of the equations of motion. An analysis of the time interval of validity for the linearized model explains the failure of the linearized and impulsive models to agree with experiment.

## 2. THE FRONT TRACKING METHOD

Front tracking is a computational method for the sharp resolution of a set of distinguished waves in a flow. It combines a standard, rectangular grid based finite difference method with a set of lower dimensional, dynamically moving grids that follow the tracked wave fronts. A general description of this method, including an outline of the structure of our computer program, is given in [10].

The numerical solution for flows in two space dimensions is represented on the union of a rectangular grid and a set of piecewise linear curves. The state at each point on the rectangular grid represents the cell average over the corresponding cell of the dual grid centered at that point. The solution at a point on a tracked front is multivalued, with values corresponding to the limits of the solution on either side of the wave. The numerical representation of the flow explicitly includes the jump discontinuities across the tracked waves and thus eliminates numerical diffusion.

Points of intersection between tracked waves, called nodes, correspond to two-dimensional interactions between wave fronts. An important example of such a node in these computations is the Mach triple point.

A global solution operator for the evaluation of the state of the flow at arbitrary locations is constructed from a front-limited triangulation of the computational grid and the tracked fronts. This triangulation is constrained so that no triangle crosses a tracked front. A side of an individual triangle in this construction is either a rectangular lattice cell boundary, or an edge on a tracked front. A corner of such a triangle is thus either a grid cell corner or a point

on a tracked front or an intersection of a tracked front with a lattice cell boundary. The states at these points serve as data for a linear interpolant of the solution into the interior of the triangle.

The representation of the solution in our front tracking code differs from the more standard triangular representations of a flow in that the tracked waves are dynamic and move with time so that the triangulation must be regenerated at each time step in the computation. The method also differs from the unstructured finite volume techniques in that the main solution is computed using a regular, rectangular grid. Subsequently the states at grid points that lie within the domain of influence of the tracked waves over the time step must be corrected to account for the presence of these waves. It is important to note that the front tracking code combines both front tracking with shock capturing, so that secondary waves (such as breaking shocks or induced slip lines) are resolved within the ability of the underlying mesh spacing and the finite difference scheme.

The propagation operator that updates the numerical solution over a single time step consists of three basic parts: propagation of the tracked wave fronts (point propagation), propagation of points of interaction between tracked waves (node propagation), and update of the states on the rectangular grid (interior solver). For the latter operation, the fronts at the beginning and end of the time step serve as internal boundaries for the regions adjoining those waves.

The first propagation phase consists of the propagation of the non-nodal points on the tracked waves. At each tracked point a local rotation of coordinates is performed that aligns the coordinate axes with the normal and tangential directions of the curve at that point. The tangent to a piecewise linear curve at a point is defined as the line through that point which is parallel to the secant through the neighboring points. Operator splitting is used to divide the front propagation into two one-dimensional units: a normal propagation step and a tangential propagation step. The normal propagation of the tracked waves is computed using a second order Riemann problem-type method as described in [5]. This operator solves a piecewise linear Cauchy problem and is similar to the van Leer [17] flux computation as used in the second order Godunov method. The tangential operator uses a one-dimensional finite difference method, which in our code is precisely the same as the interior finite difference scheme. The tangential stencil at a given point on the partially propagated wave is formed by projecting the adjacent curve states onto the tangent at that point. The interested reader is referred to references [5, 10] for a more detailed description of these operations.

An interaction between tracked waves is locally approximated by a two dimensional Riemann problem, which is defined as the Cauchy problem for initial data that is scale invariant with respect to the node position at the start of the time step, i.e. constant along rays from the node position. The numerical solution is computed using shock polar analysis. References [8, 9] contain a description of the node propagation algorithm. The primary nodes of interest in the present calculations are the Mach triple point and the diffraction node formed during the Richtmyer-Meshkov calculation by the intersection of a tracked shock wave with the material interface.

The interior finite difference scheme that computes the solution on the rectangular grid is an operator split, second order MUSCL scheme [2, 7]. Our implementation uses a five point stencil with linear reconstruction.

The rectangular grid update consists of two passes: a regular and an irregular grid update. First, the finite difference equations are solved for the the rectangular grid alone, ignoring

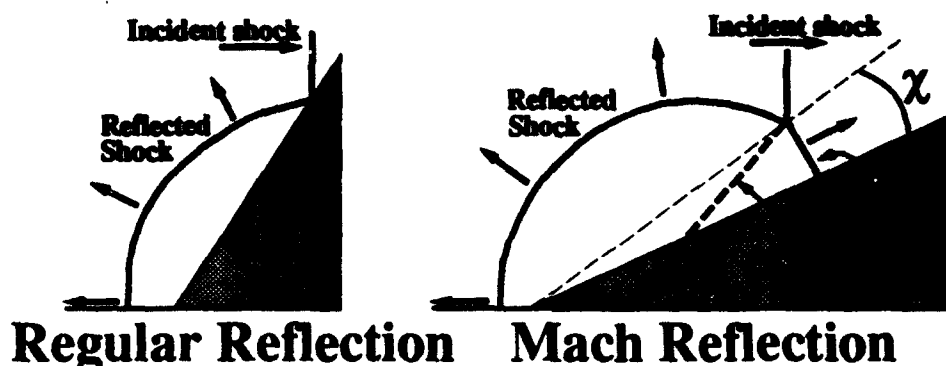


FIGURE 1. A schematic of regular and Mach reflection. Here,  $\chi$  is the Mach node trajectory angle, and  $\theta_w$  is the ramp angle. The incident shock is moving from the left to the right.

the tracked fronts. The second pass then updates the states at the rectangular grid points near the tracked fronts. If a tracked front crosses the finite difference stencil of a rectangular grid point during the time step, the states computed at that location by the first interior sweep must be discarded and recomputed to account for the presence of the tracked wave. In our implementation we use a locally modified stencil at each such point to compute the updated solution. We start with copies of the states on a five point stencil centered at the given location. We then find the tracked fronts, if any, on either side of the stencil center that are closest to the middle location. We replace any states in the stencil lying on the opposite side of the nearest front by copies of the state on the correct side of the front at the stencil crossing, as computed by linear interpolation between the tracked points on the curve. Thus the final solution never uses finite differencing across tracked waves.

The organization of the interior solver into two passes allows the bulk of the code for this part of the computation to be vectorized. Note that even though our computations used a nonvector machine (Sun Sparcstation?), this organization provides a substantial improvement in performance on vector machines and adds little additional overhead on nonvector computers.

### 3. SHOCK WAVE REFLECTION AT RAMPS

The wall reflection problems consisted of a shock wave in air (modeled as a perfect gas with  $\gamma = 1.4$ ) colliding with an oblique ramp. The gas ahead of the shock wave was at an ambient pressure of 0.3 bars and a temperature of 300° K, with an incident shock Mach number of 2.0. Two interactions were modeled with ramp angles  $\theta_w$  of 46° and 49° respectively. A computational grid of  $256 \times 256$  zones was used for both simulations, which were conducted on a Sun Microsystems Sparcstation2 with 64 megabytes of RAM. The 46° run required approximately 16 CPU hours to complete, while the 49° required about 20 CPU hours. The difference in CPU time is primarily due to the smaller value of  $\Delta t$  required in the second run. We note that our investigations have shown that a much coarser grid would resolve the basic structure, and much of the interior structure of the reflected shock bubble. We have gotten satisfactory results for the main front locations on grids as coarse as  $100 \times 100$ .

Our computation is initialized just after the shock crosses the ramp corner, when the



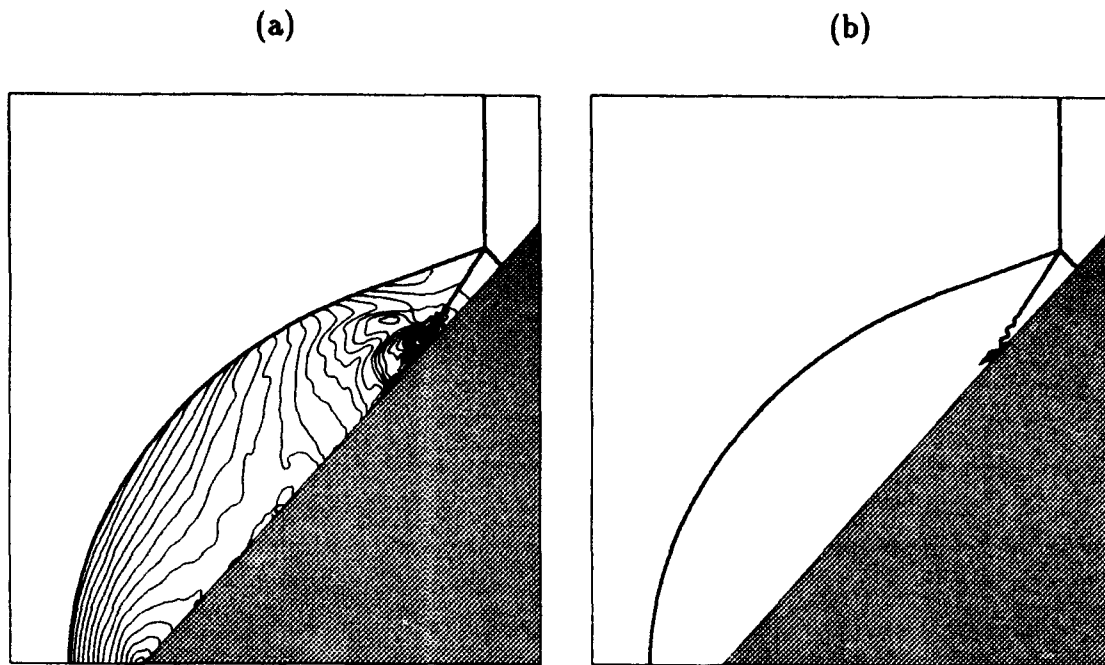


FIGURE 2. (a) Density contours from the first of the runs, using  $\theta_w = 46^\circ$ ,  $p_0 = 0.3$  bars, and  $T_0 = 300^\circ$  K. The Mach triple point trajectory angle  $\chi = 2.557^\circ$ . (b) The tracked wave fronts from the same computation.

reflected wave bubble is just a few mesh blocks in height. Tracking the complete reflected shock bubble requires an estimate of the initial geometry of this object to serve as a seed for the final computed configuration. We use the following technique to install the initial reflected waves. We assume the Mach stem is initially normal to the ramp. The Mach trajectory angle  $\chi$  is then determined algebraically by the condition that the turning angle through the incident and reflected waves in a frame that moves with the Mach triple point is the same as the turning angle through the Mach stem, i.e. the flow behind the configuration must be parallel to the slip line on both sides. The data for this system of equations consists of the ramp angle, and the incident shock data (ahead and behind states). This system is solved using an iteration on  $\chi$ , which determines the local states and wave angles about the triple point. The Mach stem and contact are installed as straight line segments from the triple point to the wall, at the computed angles. The position of the bow of the reflected shock behind the ramp is found by solving a head-on reflection Riemann problem for the state behind the incident shock. The shape of the initial reflected wave is composed of two pieces. The first is a straight piece at the triple point tangent to the computed reflected wave, and the second an ellipse from the end of this segment to the bow point, with axis on the ramp and center at the corner. It is important to note that this construction is only performed once at the beginning of the computation. The subsequent propagation of the Mach triple point and the bow node uses only local information about those points. In particular there is no restriction that the Mach stem remain straight, or that the reflected wave have any particular shape. These properties are determined dynamically by the computation. In fact, the shapes of the waves at later times appear to be independent of any reasonable initial configuration. Figure 1 shows a schematic of the basic geometry of the reflection.

Figures 2 and 3 show the results of our computations of the two simulations outlined above. Figures 2a and 3a show density contours, while figures 2b and 3b show only the tracked wave

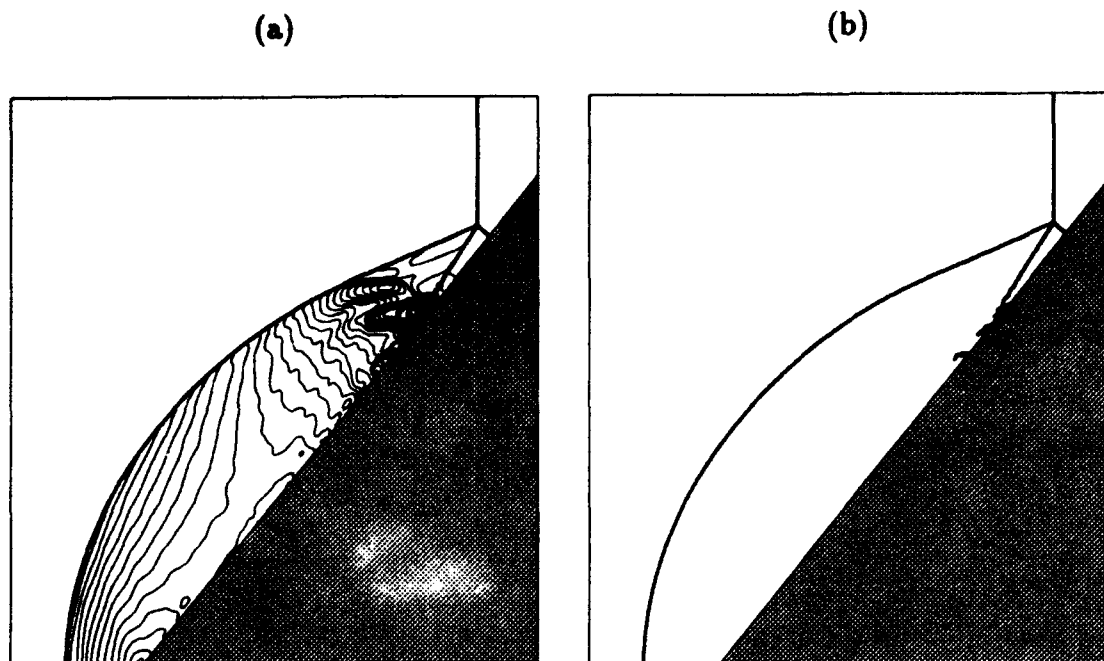


FIGURE 3. (a) Density contours for the second of the runs. This computation differs from the first only in that  $\theta_w = 49^\circ$ . Here the Mach triple point trajectory angle  $\chi = 1.6^\circ$ . (b) The tracked wave fronts from the same computation.

fronts. The density difference of adjacent contours is 4% of the density of the unshocked air. We see very sharp resolution of the tracked waves, and in particular at the triple point. This resolution becomes even more important as we move closer to the bifurcation point to regular reflection, which we will discuss more fully in the context of the next series of simulations. Let us just say here that our code supports the von Neumann criterion over the detachment criterion as the location of this point, and these two runs appear to bear that out, since  $\theta_N = 53.36^\circ$ , and  $\theta_e = 50.898^\circ$ . We are relatively close to  $\theta_e$  in Figure 3, but still have a distinct Mach reflection structure with a sizeable Mach stem.

There is a considerable amount of activity in the flow at the point where the slip line induced from the Mach triple point reaches the wall. The slip boundary conditions at the wall require that the flow there be parallel to it, and hence there is a large gradient in the velocity as the flow adjusts to the obstacle. Tracking the slip line reduces the amount of numerical spreading of this wave at the wall, which in turn enhances the resolution of the flow about the triple point by preventing its contamination by the transient waves produced at the boundary.

The density contours of figures 2a and 3a show that our computation is doing a good job of reducing the effect of wall heating at the ramp boundary. Most of the contours are relatively smooth going into the boundary, which is consistent with the inviscid model used for these computations. We made no attempt to model the boundary layer effects that are present in a real experiment.

Another feature of our code is the ability to measure various physical quantities very precisely with respect to our computations. For example, we know the exact states around the triple point or at the base of the Mach stem by simply reading them from the curve data structures in the computation.

In a separate series of runs, we validated our computations of regular Mach reflections

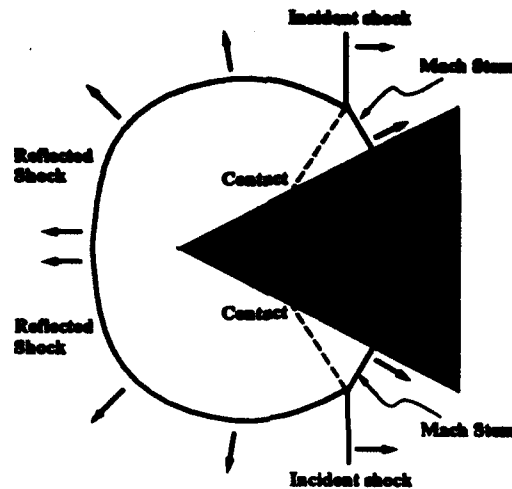


FIGURE 4. A schematic representation of the experimental apparatus used to produce Mach reflections. This configuration produces a pair of symmetric reflections as the incoming shock passes the apex of the wedge.

by comparing them with experiments performed at the University of Sydney by Henderson and Virgona [11]. Considerable effort was made to simplify the experiments in order to facilitate comparison with the numerical results. For example, the Mach reflections were generated in argon ( $\gamma = 1.667$ ) so as to eliminate vibrational non-equilibrium, dissociation, and chemical reactions. The strength of the incident shock  $i$  was sufficiently large to ensure that the flow downstream of the reflected shock  $r$  was supersonic, but not so strong as to ionize the argon. More precisely, the average strength of  $i$  used in the experiments, or rather the average inverse strength  $\xi_i \equiv p_0/p_1$  was  $\xi_i = 0.1534$ , corresponding to an incident Mach number of 2.327. The ahead state pressure and temperature were  $p_0 = 14.1 \pm 3.0$  kPa and  $T_0 = 293.15 \pm 4.0^\circ$  K. The Mach reflections were generated by diffracting incident shocks over a series of symmetrical wedges of different apex semi-angles  $\theta_w$  (Figure 4). This design eliminated shock-boundary layer interaction at the apex of every wedge. The more conventional concave corner model (Figure 1) undergoes significant shock-boundary layer interaction as the reflected shock  $r$  sweeps over the lower wall. The newer model eliminates this effect, and is mathematically equivalent to the model in Figure 1 in the inviscid case. Considerable development work was done on the design of the boundary layer spill slots on the side walls of the shock tube. The objective was to minimize, as far as practical, the shock boundary layer on the side walls of the tube.

A graph of Mach triple point trajectory,  $\chi$ , versus the ramp angle,  $\theta_w$ , is shown in Figure 5a. These figures show the experimentally measured value of  $\chi$  together with the values computed by our front tracking code, as well as those computed by Colella [11] using a highly resolved shock capturing scheme. This measurement was attractive because  $\chi$  cannot be computed from the shock polar analysis used to compute the local configuration at the node. Its value is entirely a result of the interaction of the numerics in our code, and its experimental value is particularly robust. We have also found the Mach number behind the base of the Mach stem to be a very useful measurement, for exactly the same reasons.

As can be seen from the picture, our results agree very well with both the experimental and the shock capturing results. We note that both sets of computational results are on the high side of the experimental values. This difference has been attributed to boundary layer

## Wall Angle vs Mach Node Trajectory

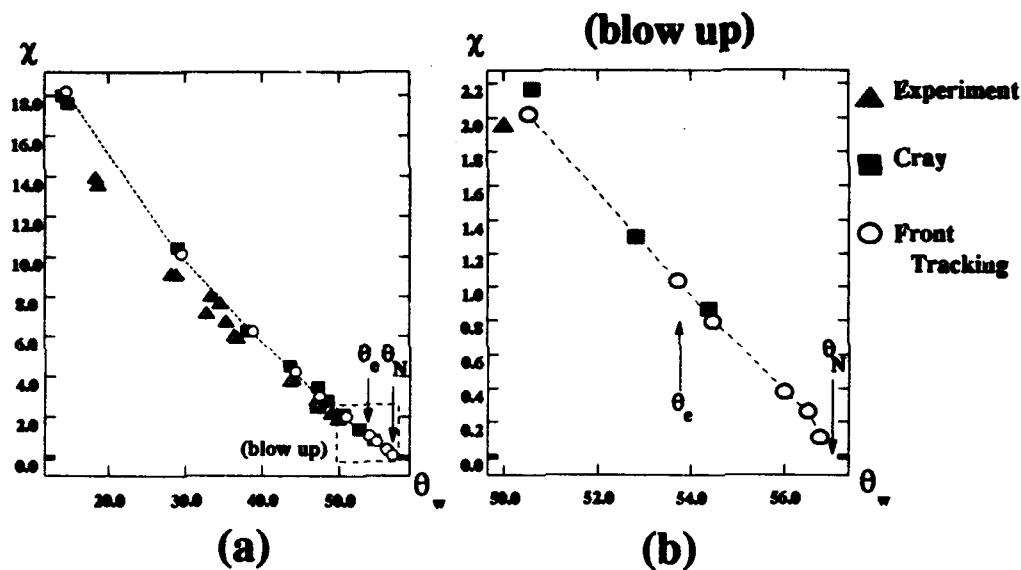


FIGURE 5. A comparison of the Mach triple point trajectory for experiments, fine grid shock capturing computations, and front tracking. We note that both numerical methods are in substantial agreement with each other, and are close to the experimental measurements. Front tracking provides approximately the same value for  $\chi$  as the shock capturing code, using only a fraction of the grid resolution.

effects.

An application of important interest is the transition conditions between regular and Mach reflection at a wall. It is well known that for certain flow regimes there is an overlap between the regions in phase space where regular and Mach reflections are possible. Both experimental and computational investigations have shown that the boundary layer at the wall plays an important role in the process that selects the type of wave produced by the wall reflection. In order to quantify the effect of the wall boundary layer it is important to understand the inviscid limit of the solution where this boundary layer is absent. It is in the computation of this inviscid limit that a major strength of the front tracking method is revealed. Since we are explicitly tracking the most singular parts of the calculation, we can make very precise statements about exactly where a given discontinuity is located; there is no numerical diffusion of the fronts. This allows us to perform computations near the transition to regular reflection, yielding structure that is not resolvable in either experiment or standard shock capturing codes. We can resolve the full Mach triple point configuration for angles  $\chi$  as small as  $0.1^\circ$  (Figure 5b). In such simulations the Mach stem is less than a grid block long. By contrast, shock capturing codes generally lose the resolution of the Mach triple point when the length of Mach stem is less than two or three grid blocks. This loss of resolution is due to the presence of a numerical boundary layer at the wall. This regime is also difficult to approach experimentally due to the real viscous boundary layer at wall. Currently, front tracking appears to be the only method that can conduct numerical simulations of inviscid wall reflections to within a small fraction of a degree of the mechanical equilibrium condition.

The resolution for the front tracking runs was achieved on grids which are much coarser than those used in standard finite difference simulations of this problem. Most of our grids

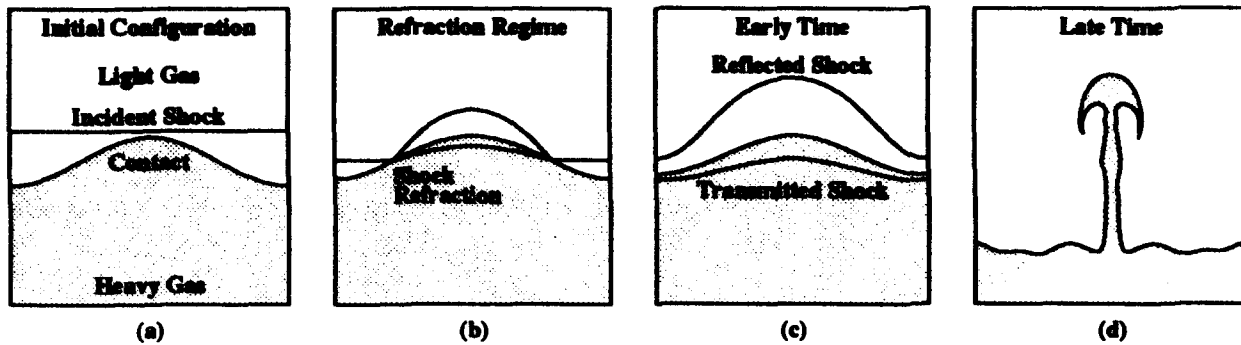


FIGURE 6. A schematic representation of the geometry of the Richtmyer-Meshkov instability modeled in this paper. The interaction consists of the collision of a shock wave with a material interface. The refraction of the shock by the interface produces reflected and transmitted waves. The instability consists of the growth of perturbations of the material interface with time.

were  $100 \times 100$ , up to about  $150 \times 150$  close to the transition point. This also gives a commensurate savings in time – the runs took between one and four hours.

It should also be pointed out that for the region between  $\theta_e$  and  $\theta_N$ , our code can simulate either regular or Mach reflection – both are theoretically possible in this region. However, based on the shape of the curve outside this region (see Figure 5b), our results definitely seem to converge to the point  $\theta_N$ , and there is no reason to expect a discontinuity in the curve at  $\theta_e$ . We feel that this is a very strong statement that in this parameter regime, the bifurcation from Mach Reflection to regular reflection takes place at the mechanical equilibrium condition and not at the detachment point.

#### 4. NUMERICAL SIMULATION OF THE RICHTMYER-MESHKOV INSTABILITY

We focus on the simplest case of the shock tube experiments of the Richtmyer-Meshkov instability where a sine shaped material interface is accelerated by a single shock wave, as in the experiments of Meshkov [12], Benjamin [3, 4], and others. The general configuration of the computation and experiments is shown in Fig. 6. A thin membrane was used in the experiments to separate the two gases at the material interface. Quantitative agreement was achieved between our computational results and the experimental measurements of Benjamin [4] for the rate of growth of a shocked air-SF<sub>6</sub> interface. The collision results in a transmitted shock and a reflected wave that can be either a shock or a rarefaction depending on the values of the fluid parameters. The experiments considered in this paper are of the reflected shock type. Viscosity and heat conduction are negligible here, so the fluid motion is described by the Euler equations.

The impulsive model proposed by Richtmyer [15] is commonly used to estimate the growth rate of a shock accelerated interface. This model is derived by assuming that the shock acceleration can be treated as being impulsive, and that the flow is nearly incompressible once the shock wave has passed through the material interface. It is also assumed that the flow is observed in a frame where the average position of the material interface is at rest, and the position,  $y(x, t)$ , of the material interface at time  $t$  can be given by  $y(x, t) = a(t) \sin kx$ , where  $k$  is the wave number of the perturbation. Richtmyer's formula gives the growth rate

of  $a(t)$  as

$$\dot{a}(t) = k\Delta u \frac{\rho_1 - \rho_2}{\rho_1 + \rho_2} a(0+), \quad (4.1)$$

where  $\Delta u$  is the difference between the shocked and unshocked mean interface velocities, the  $\rho_i$  are the post-shocked densities on the two sides of the interface (the incident shock moves from material "2" to material "1"), and  $a(0+)$  is the perturbation amplitude immediately after the collision of the shock with the material interface. This formula implicitly assumes the initial preshocked amplitude,  $a(0-)$ , is small compared to the wavelength.

Given that  $a(0-)$  is small so that  $ka(0-) \ll 1$ , a more exact calculation of the amplitude growth rate can be made. The Euler equations are linearized around the solution of a one dimensional Riemann problem defined by the head-on collision of a planar shock with a zero amplitude (planar) material interface, using the initial amplitude of the sinusoidal perturbation as a small expansion parameter. The result of the linearization is a system of partial differential equations in one spatial dimension with associated boundary conditions. This system can be solved numerically for the growth rate of the perturbed interface. This approach, following Richtmyer [15], has recently been generalized to include reflected rarefactions as well as reflected shocks [18]. Simple order of magnitude estimates limit the validity of the linearized equations to the dimensionless time interval

$$t_{*min} \equiv ka(0-) \ll t_* \ll 1/[ka(0-)] \equiv t_{*max}. \quad (4.2)$$

Here the dimensionless time  $t_* = kc_0 M_0 t$ , where  $M_0$  is the incident shock Mach number, and  $c_0$  is the sound speed of the fluid ahead of the incident shock. The limits  $t_{*min}$  and  $t_{*max}$  represent respectively the transit time of the incident shock through the perturbed interface and the time required for the perturbation to grow to unit amplitude. Necessarily, these time limits apply to the derivation of the impulsive model as well, since it is an approximation to the linear theory. Recent systematic comparisons of the impulsive model and the linear theory have revealed both regions of agreement and of disagreement in parameter space [18].

We compared our simulations of a singly shocked air-SF<sub>6</sub> interface to the experiments of Benjamin [4]. The material interface is accelerated by a shock wave with Mach number 1.2 moving from air into SF<sub>6</sub>. The initial amplitude,  $a(0-)$ , was 0.00637 times the period of the sinusoidal perturbation. For these experiments,  $t_{*max} \approx 2.5$ , while the observational time interval is  $15 \leq t_{*observational} \leq 50$ . The observational times and the validity of the linear theory fail to overlap by a factor of about 6. We conclude that the linear theory has no relationship to this experiment.

Fig. 7 shows plots of the amplitude and amplitude growth rate of the material interface as obtained from experiment, the front tracking simulation, the linearized theory, and Richtmyer's impulsive model. The time axis in these figures is shifted so that  $t = 0$  corresponds to the time at which the shock wave has completed its refraction through the interface.

As can be seen from these figures the front tracking results are in substantial agreement with the experimental results in the sense that the growth rate derived by a least squares analysis of the amplitude data, 8.14 m/s, is within the experimental range of 7.9 m/s  $\pm$  10%. Note that for late (i.e., experimentally observed) times the linearized theory and the impulsive model growth rates are a factor of two larger than those found in experiment or in our simulation. This may be due to the fact that this particular configuration has a relatively large initial amplitude and quickly leaves the region of validity of the linearized theory and impulsive model. The displacement of the experimental curve with respect to

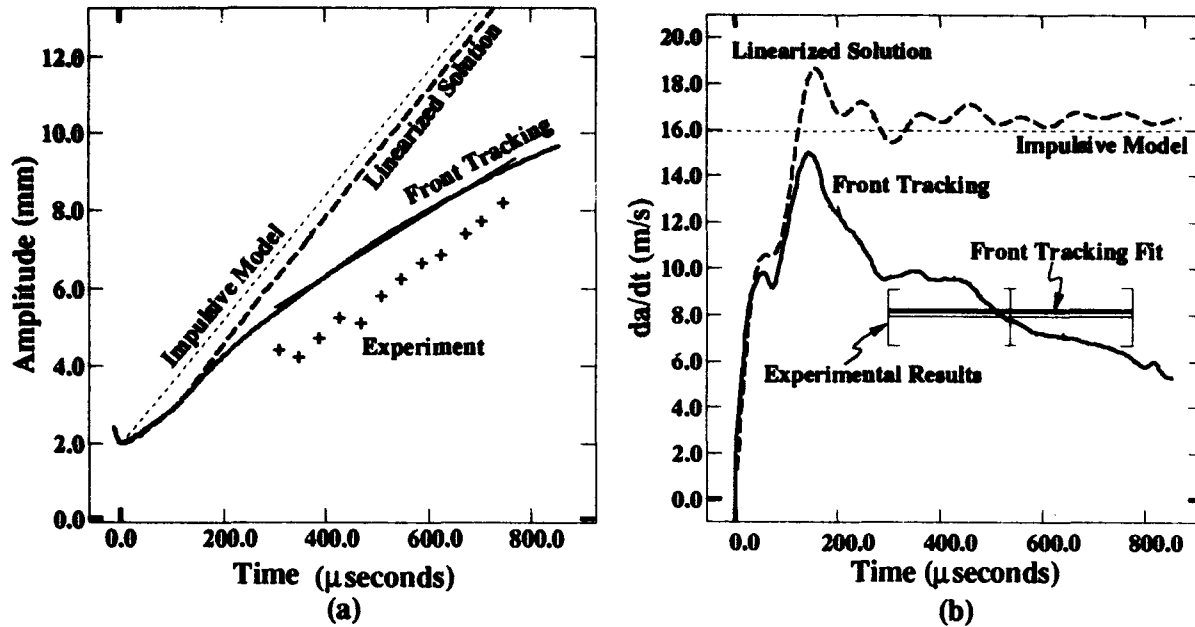


FIGURE 7. Perturbation amplitude,  $a(t)$ , and amplitude growth rate,  $\dot{a}(t)$ , of a shocked air-SF<sub>6</sub> interface. This graph compares the results of experimental averages, front tracking simulation, linear theory and Richtmyer's impulsive model. Also shown are results of a least squares fit to the front tracking amplitude data over the period of experimental observation. The plus marks (+) show the results of one particular experiment, while the experimental growth rate represents an average over several experiments.

the front tracking curve is possibly due to membrane effects, i.e. the material strength of the membrane or the influence of its fragmentation may effect the fluid flow.

The front tracking results indicate a decay in amplitude growth rates while Benjamin [4] finds a fairly constant growth rate during the measurement period. Other experiments, however, have shown a decaying growth rate [13, 1]. The figures shown in this paper used a resolution of 125 zones per wavelength, and mesh refinement studies in the range of 125–208 zones per wavelength showed very little change in the amplitude growth rate. We also tested our simulation against changes in other numerical parameters and found that the value of  $\dot{a}(t)$  was insensitive to these changes. We conclude that this decay is a real effect and not due simply to numerical dissipation as has been suggested [4].

A further validation of the nonlinear simulations can be accomplished by comparison to the small amplitude theory (Fig. 8). This serves both to determine the range of validity of the linear theory and to validate the solution of the full Euler equations at small amplitudes. As can be seen in Fig. 8, the front tracking calculation is converging to the linear result as we reduce the amplitude. We note that the interval of convergence of the nonlinear simulations to the linear theory appears to be finite. This is in contrast to formula 4.2 which suggests that the domain of validity of the linearized equations should increase with decreasing initial amplitude. This point deserves further study.

Of interest is the question of why our results agree with experiment while results found

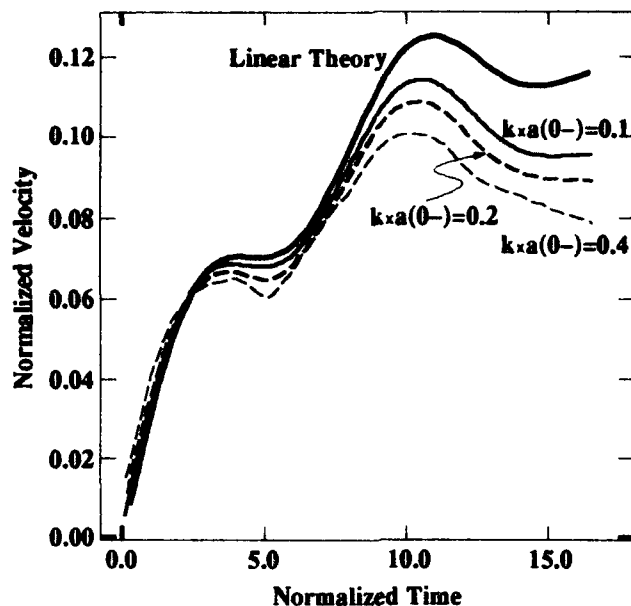


FIGURE 8. The convergence of the nonlinear simulations to the linearized solution for small amplitudes. A comparison of three separate calculations of the normalized perturbation growth rate,  $\dot{a}(t)/[kc_0M_0a(0-)]$ , of a shocked air-SF<sub>6</sub> interface with three different initial amplitudes where  $k$  is the wave number,  $c_0$  is the sound speed ahead of the incident shock, and  $M_0$  is the incident shock Mach number. The horizontal axis is in dimensionless time units  $kc_0M_0t$ .

through other numerical methods do not. Prior disagreement between the growth rates measured in experiments and those predicted by numerical simulation has led to the suggestion that mass diffusion and membrane effects may have an important role in the behavior of the interface instabilities. Our work does not exclude this possibility, but the agreement of our computations with experiment suggests that a proper numerical resolution of the material interface is essential to obtain agreement with experiment, and also that if other effects are important, they may be offsetting one another. It is also clear that there is still much to learn about the highly nonlinear aspects of the Richtmyer-Meshkov instability. These effects include the possible coupling between nonlinear modes, and their study will require experiments on singly shocked interfaces as well as computations with random interfaces which have been run to late times. Similarly, understanding the effects of reshocking remains an important theoretical challenge. For the single mode case, a systematic study of mass diffusion, membrane effects, and a detailed comparison to earlier calculations of others would be desirable.

#### ACKNOWLEDGMENTS

We thank James Glimm for his encouragement and guidance.

#### REFERENCES

1. A. N. Aleshin, E. G. Gamalli, S. G. Zaitsev, E. V. Lazareva, I. G. Lebo, and V. B. Rozanov. Nonlinear and transitional stages in the onset of the Richtmyer-Meshkov instability. *Sov. Tech. Phys. Lett.*, 14:466-468, 1988.



## SURFACE INSTABILITIES AND WAVE INTERACTIONS

2. J. Bell, P. Colella, and J. Trangenstein. Higher order Godunov methods for general systems of hyperbolic conservation laws. *J. Comput. Phys.*, 82:362-397, 1989.
3. R. Benjamin. Experimental observations of shock stability and shock induced turbulence. In W.P. Dannevik, A.C. Buckingham, and C.E. Leith, editors, *Advances in Compressible Turbulent Mixing*, pages 341-348. National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd. Springfield VA 22161, 1992.
4. R. Benjamin, D. Besnard, and J. Haas. Shock and reshock of an unstable interface. LANL report LA-UR 92-1185, Los Alamos National Laboratory, 1993.
5. I-L. Chern, J. Glimm, O. McBryan, B. Plohr, and S. Yaniv. Front tracking for gas dynamics. *J. Comput. Phys.*, 62:83-110, 1986.
6. L. D. Cloutman and M. F. Wehner. Numerical simulation of Richtmyer-Meshkov instabilities. *Phys. Fluids A*, 4:1821-1830, 1992.
7. P. Colella. A direct Eulerian MUSCL scheme for gas dynamics. *SIAM Journal on Computing*, 6:104, 1985.
8. J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv. Front tracking and two dimensional Riemann problems. *Adv. Appl. Math.*, 6:259-290, 1985.
9. J. Grove. The interaction of shock waves with fluid interfaces. *Adv. Appl. Math.*, 10:201-227, 1989.
10. J. Grove. Applications of front tracking to the simulation of shock refractions and unstable mixing. In *J. Appl. Num. Math.*, 1993. to appear.
11. L. F. Henderson, P. Colella, and R. J. Virgona. Strong shock reflection in pseudo-stationary flow, 1993.
12. E. E. Meshkov. Instability of a shock wave accelerated interface between two gases. *NASA Tech. Trans.*, F-13:074, 1970.
13. E. E. Meshkov. Instability of shock-accelerated interface between two media. In W.P. Dannevik, A.C. Buckingham, and C.E. Leith, editors, *Advances in Compressible Turbulent Mixing*, pages 473-503. National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Rd. Springfield VA 22161, 1992.
14. K. A. Meyer and P. J. Blewett. Numerical investigation of the stability of a shock-accelerated interface between two fluids. *Phys. Fluids*, 15:753-759, 1972.
15. R. D. Richtmyer. Taylor instability in shock acceleration of compressible fluids. *Comm. Pure Appl. Math.*, 13:297-319, 1960.
16. V. Rupert. Shock-interface interaction: current research on the Richtmyer-Meshkov problem. In K. Takayama, editor, *Shock Waves, proceedings of the 18th international symposium on shocks waves*. Springer-Verlag, New York, 1992.
17. B. van Leer. Towards the ultimate conservative difference scheme: V. a second order sequel to Godunov's method. *J. Comp. Phys.*, 32:101-136, 1979.
18. Y. Yang, Q. Zhang, and D. H. Sharp. Small amplitude theory of Richtmyer-Meshkov instability. Report No. SUNYSB-AMS-93-08, State Univ. of New York at Stony Brook, 1993. LANL report LA-UR 93-2535.

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS, STATE UNIVERSITY OF NEW YORK AT STONY BROOK, STONY BROOK, NY 11794-3600

E-mail address: boston@ams.sunysb.edu, grove@ams.sunysb.edu, holmes@ams.sunysb.edu, henderso@ams.sunysb.edu, ymy@ams.sunysb.edu, zhang@ams.sunysb.edu

COMPLEX SYSTEMS GROUP, THEORETICAL DIVISION, LOS ALAMOS NATIONAL LABORATORY, LOS ALAMOS, NM 87545

E-mail address: dhs@t13.lanl.gov

# STABLE COMPACT SCHEMES FOR SHOCK CALCULATIONS

Bernardo Cockburn <sup>1</sup>  
School of Mathematics  
University of Minnesota  
Minneapolis, MN 55455

and

Chi-Wang Shu <sup>2</sup>  
Division of Applied Mathematics  
Brown University  
Providence, RI 02912

**ABSTRACT.** We discuss the applications of high order compact finite difference methods for shock calculations. Nonlinear stability is achieved through the definition of a local mean which serves as a reference for introducing a local flux limiting to control spurious numerical oscillations while keeping the formal accuracy of the scheme. For scalar conservation laws, the resulting schemes can be proven total variation stable in one space dimension and maximum norm stable in multi space dimensions. Numerical examples are shown to verify accuracy and stability of such schemes for problems containing shocks.

## 1 Introduction

Compact schemes are methods where the derivatives are approximated not by polynomial operators but by rational function operators on the discrete solutions. We are interested in solving a hyperbolic conservation law

$$\begin{aligned}u_t + f(u)_x + g(u)_y &= 0 \\ u(x, y, 0) &= u^0(x, y)\end{aligned}\tag{1.1}$$

using compact schemes. In the semi-discrete form, a compact scheme for solving (1.1) can be written as

$$\frac{du_{ij}}{dt} = -\frac{1}{\Delta x}(A_x^{-1}B_x f(u))_{ij} - \frac{1}{\Delta y}(A_y^{-1}B_y g(u))_{ij} \equiv L(u)_{ij}\tag{1.2}$$

where  $A$  and  $B$  are both *local*, one dimensional operators. The subscript  $x$  or  $y$  indicates that the operator is applied in the  $x$  or  $y$  direction.

As examples, a fourth order central compact scheme is given by (1.2) with

$$(Av)_i = \frac{1}{6}(v_{i-1} + 4v_i + v_{i+1})$$

---

<sup>1</sup>Research partly supported by the National Science Foundation grant DSM-9103997 and by the Minnesota Supercomputer Institute.

<sup>2</sup>Research supported by ARO grant DAAL03-91-G-0123, NSF grant DMS-9211820, NASA Langley grant NAG1-1145 and AFOSR grant 93-0090.

$$(Bv)_i = \frac{1}{2}(v_{i+1} - v_{i-1}) \quad (1.3)$$

and two third order upwind compact schemes are given by

$$\begin{aligned} (Av)_i &= \frac{1}{3}(-v_{i-1} + 5v_i - v_{i+1}) \\ (Bv)_i &= \frac{1}{2}(3v_i - 4v_{i-1} + v_{i-2}) \end{aligned} \quad (1.4)$$

and

$$\begin{aligned} (Av)_i &= \frac{1}{3}(-v_{i-1} + 5v_i - v_{i+1}) \\ (Bv)_i &= \frac{1}{2}(-v_{i+2} + 4v_{i+1} - 3v_i) \end{aligned} \quad (1.5)$$

respectively, depending upon the wind direction. Notice that (1.4) and (1.5) have the same implicit part  $A$  which is symmetric. This fact will be used later in Section 2 to define our local means.

The cost of compact schemes, regardless of the number of space dimensions, involves only inversion of the narrowly banded (usually tridiagonal) matrix  $A$ , hence is comparable to explicit methods. This is notably different from other implicit methods such as the continuous Galerkin finite element methods in multi space dimensions, even if they are similar in one space dimension.

The advantages of compact schemes over traditional finite difference methods include the relatively high order of accuracy using a compact stencil (for example, the fourth order scheme (1.3), when discretized in time using Euler forward, uses only a three point stencil in each time level), a better (linear) stability, a better resolution for high frequency waves [13], and usually fewer boundary points to handle. In recent years compact schemes have attracted considerable attention in various fields such as the direct numerical simulations of turbulence. We refer the readers to [13], [19], [12], [3], [1], and [2] for more details. The recent paper [13] discusses in detail wave resolution, phase errors and other issues related to compact schemes and is a good reference.

We are interested in applying compact schemes for shock calculations. Just like any other linear schemes (schemes which are linear when applied to linear equations), compact schemes usually demonstrate nonlinear instability when applied to discontinuous data. We follow the TVD (total variation diminishing) ideas in [9], [14] and try to define a suitable nonlinear local limiting to avoid spurious oscillations while keeping the formal accuracy of the scheme. Notice that the compact scheme, just like any implicit scheme, is global. That is, the approximation to  $f(u)_x$  at  $x = x_i$  involves  $u_k$  along the whole line due to the tridiagonal inversion  $A^{-1}$ . Our main idea is to define a *local* mean, and to use it as a reference for introducing a local limiting. In Section 2 we introduce the limiting for one space dimension and obtain total variation stability. In Section 3 we introduce the limiting for multi space dimensions and obtain maximum norm stability. In Section 4 we present numerical examples. In most cases we will only state the theoretical results without proof. We refer the readers to [6] for details.

In this paper, we use the total variation diminishing (TVD) Runge-Kutta type time discretization, introduced in [17], [15], to discretize the ODE in the method-of-lines formulation (1.2). In the third order case, the time discretization is

$$\begin{aligned} u^{(1)} &= u^n + \Delta t L(u^n) \\ u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}u^{(1)} + \frac{1}{4}\Delta t L(u^{(1)}) \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}u^{(2)} + \frac{2}{3}\Delta t L(u^{(2)}) \end{aligned} \quad (1.6)$$

Only third order results will be shown, although schemes with other orders of accuracy are also tested.

These special Runge-Kutta type time discretizations are labelled TVD because it can be proven that, under suitable restrictions on the time step  $\Delta t$  (the CFL condition), the full discretization (1.6) is TVD, or stable under another norm, for example the  $L_\infty$  norm, if the first order Euler forward time discretization for (1.2):

$$u^{n+1} = u^n + \Delta t L(u^n) \quad (1.7)$$

is TVD or stable under the other norm. For details, see [17] and [15].

We thus only need to consider the Euler forward scheme (1.7) for stability analysis in the subsequent sections.

## 2 One Space Dimension

In one space dimension, equation (1.1) becomes

$$\begin{aligned} u_t + f(u)_x &= 0 \\ u(x, 0) &= u^0(x) \end{aligned} \quad (2.1)$$

the scheme (1.2) is

$$\frac{du_i}{dt} = -\frac{1}{\Delta x} (A^{-1} B f(u))_i \equiv L(u)_i \quad (2.2)$$

and the Euler forward time discretization (1.7) becomes

$$u_i^{n+1} = u_i^n + \Delta t L(u^n)_i \quad (2.3)$$

Scheme (2.3) can be easily written into a conservation form

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} (h_{i+\frac{1}{2}}^n - h_{i-\frac{1}{2}}^n) \quad (2.4)$$

suitable for shock calculations. However, the numerical flux  $h_{i+\frac{1}{2}}^n$  is not a local function of  $u_k^n$  due to the tridiagonal inversion  $A^{-1}$ . If we define

$$\bar{u}_i \equiv (Au)_i \quad (2.5)$$

then scheme (2.3) can be left-multiplied by  $A$  to become

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{\Delta x} (Bf(u^n))_i \quad (2.6)$$

which, when written into a conservation form,

$$\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{\Delta x} (\hat{f}_{i+\frac{1}{2}}^n - \hat{f}_{i-\frac{1}{2}}^n) \quad (2.7)$$

involves a numerical flux  $\hat{f}_{i+\frac{1}{2}}^n$  which is a local function of  $u_k^n$ . For example,

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2} (f(u_{i+1}) + f(u_i)) \quad (2.8)$$

for the fourth order central scheme (1.3); and

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2} (3f(u_i) - f(u_{i-1})) \quad (2.9)$$

and

$$\hat{f}_{i+\frac{1}{2}} = \frac{1}{2} (-f(u_{i+2}) + 3f(u_{i+1})) \quad (2.10)$$

for the two third order upwind schemes (1.4) and (1.5), respectively. Notice that scheme (2.7) resembles a cell-averaged (finite volume) scheme [11]. The  $\bar{u}_i$  in (2.5), just like a cell average, is a local mean of  $u$ , defined by  $Au$  in (1.3) through (1.5). Since the computation of the flux  $\hat{f}_{i+\frac{1}{2}}$  in (2.7) involves the values of  $u$ , a "reconstruction" from  $\bar{u}$  to  $u$

$$u_i = (A^{-1}\bar{u})_i \quad (2.11)$$

is needed. This reconstruction is global.

It is now rather straightforward to define the limiting. We first write

$$f(u) = f^+(u) + f^-(u) \quad (2.12)$$

with the requirement that

$$\frac{\partial f^+(u)}{\partial u} \geq 0, \quad \frac{\partial f^-(u)}{\partial u} \leq 0 \quad (2.13)$$

The purpose of this flux splitting is for easier upwinding at later stages. The simplest such splitting is due to Lax-Friedrichs

$$f^\pm(u) = \frac{1}{2} (f(u) \pm \alpha u), \quad \alpha = \max_u |f'(u)| \quad (2.14)$$

where the maximum is taken over the range of  $u^0(x)$ . We then write the flux  $\hat{f}_{i+\frac{1}{2}}$  in (2.7) also as

$$\hat{f}_{i+\frac{1}{2}} = \hat{f}_{i+\frac{1}{2}}^+ + \hat{f}_{i+\frac{1}{2}}^- \quad (2.15)$$

where  $\hat{f}_{i+\frac{1}{2}}^\pm$  are obtained by putting superscripts  $\pm$  in (2.8) through (2.10).

Next we define

$$d\hat{f}_{i+\frac{1}{2}}^+ = \hat{f}_{i+\frac{1}{2}}^+ - f^+(\bar{u}_i); \quad d\hat{f}_{i+\frac{1}{2}}^- = f^-(\bar{u}_{i+1}) - \hat{f}_{i+\frac{1}{2}}^- \quad (2.16)$$

Here  $d\hat{f}_{i+\frac{1}{2}}^\pm$  are the differences between the numerical fluxes  $\hat{f}_{i+\frac{1}{2}}^\pm$  and the first order, upwind fluxes  $f^+(\bar{u}_i)$  and  $f^-(\bar{u}_{i+1})$ . These differences are subject to limiting for nonlinear stability. We define the limiting by

$$\begin{aligned} d\hat{f}_{i+\frac{1}{2}}^{+(m)} &= m\left(d\hat{f}_{i+\frac{1}{2}}^+, \Delta_+ f^+(\bar{u}_i), \Delta_+ f^+(\bar{u}_{i-1})\right) \\ d\hat{f}_{i+\frac{1}{2}}^{-(m)} &= m\left(d\hat{f}_{i+\frac{1}{2}}^-, \Delta_+ f^-(\bar{u}_i), \Delta_+ f^-(\bar{u}_{i+1})\right) \end{aligned} \quad (2.17)$$

where  $\Delta_+ v_i \equiv v_{i+1} - v_i$  is the usual forward difference operator, and the (now standard) *minmod* function  $m$  is defined by

$$m(a_1, \dots, a_k) = \begin{cases} s \min_{1 \leq i \leq k} |a_i|, & \text{if } \text{sign}(a_1) = \dots = \text{sign}(a_k) = s \\ 0, & \text{otherwise} \end{cases} \quad (2.18)$$

See, e.g., [9]. Notice that the limiting defined in (2.17) is upwind biased.

The limited numerical fluxes are then defined by

$$\hat{f}_{i+\frac{1}{2}}^{+(m)} = f^+(\bar{u}_i) + d\hat{f}_{i+\frac{1}{2}}^{+(m)}; \quad \hat{f}_{i+\frac{1}{2}}^{-(m)} = f^-(\bar{u}_{i+1}) - d\hat{f}_{i+\frac{1}{2}}^{-(m)} \quad (2.19)$$

and

$$\hat{f}_{i+\frac{1}{2}}^{(m)} = \hat{f}_{i+\frac{1}{2}}^{+(m)} + \hat{f}_{i+\frac{1}{2}}^{-(m)} \quad (2.20)$$

If we define the total variation of the mean  $\bar{u}$  by

$$TV(\bar{u}) = \sum_i |\bar{u}_{i+1} - \bar{u}_i| \quad (2.21)$$

we have the following

**Proposition 2.1:** Scheme (2.7) with the flux (2.20) is TVDM (total variation diminishing in the means):

$$TV(\bar{u}^{n+1}) \leq TV(\bar{u}^n) \quad (2.22)$$

under the CFL condition

$$\max_{\min_i, \bar{u}_i^n \leq u \leq \max_i, \bar{u}_i^n} (f^{+'}(u) - f^{-'}(u)) \frac{\Delta t}{\Delta x} \leq \frac{1}{2} \quad (2.23)$$

□

The limiting defined in (2.17) is just one of many possibilities. See [18] for a comprehensive discussion of limiters.

The total variation stability for  $u$  itself is based upon the previous proposition and the following lemma:

**Lemma 2.2:** If there are two numbers  $0 \leq \delta < 1$  and  $\alpha > 0$ , which are independent of  $N$ , such that the  $N \times N$  matrix  $A = (a_{ij})$  satisfies:

$$\max_{1 \leq j \leq N} \frac{1}{|a_{jj}|} \leq \alpha, \quad \text{and} \quad \sum_{\substack{i=1 \\ i \neq j}}^N |a_{ij}| \leq \delta |a_{jj}|, \quad j = 1, \dots, N \quad (2.24)$$

(strongly diagonally dominance for the transpose of  $A$ ), then the  $L_1$  norm of  $A^{-1}$  is bounded independent of  $N$ :

$$\|A^{-1}\|_{L_1} \leq \frac{\alpha}{1 - \delta} \quad (2.25)$$

□

For most compact methods, the matrix  $A$  satisfies the condition (2.24) for Lemma 2.2. For example, in the schemes defined by (1.3), (1.4) and (1.5),  $A$  satisfies the condition (2.24) with  $\delta = \frac{1}{2}$ ,  $\alpha = 6$ ;  $\delta = \frac{2}{5}$ ,  $\alpha = 3$  and  $\delta = \frac{2}{5}$ ,  $\alpha = 3$ , respectively. For such compact schemes, we have the total variation stability for  $u$ :

**Proposition 2.3:** If a compact scheme (2.7) satisfies the conditions in Proposition 2.1 and Lemma 2.2, then it is TVB (total variation bounded). That is,

$$TV(u^n) = \sum_i |u_{i+1}^n - u_i^n| \leq C \quad (2.26)$$

for all  $n \geq 0$  and  $\Delta t > 0$ . Here  $C$  is a constant independent of  $n$  and  $\Delta t$ .

□

This Proposition guarantees convergence of at least a subsequence of the numerical solution.

We now discuss whether the limiting defined in (2.17) maintains the *formal* accuracy of the compact schemes in smooth regions of the solution. For this we need the following

**Assumption 2.4:**

$$\bar{u}_i = (Au)_i = u_i + O(\Delta x^2) \quad (2.27)$$

for all  $u \in C^2$ .

□

This Assumption is satisfied by any compact scheme with a *symmetric*  $A$ , for example all those listed in (1.3) through (1.5).

Under Assumption 2.4, it is easy to verify, by simple Taylor expansions, that

$$\begin{aligned} \Delta_+ f^\pm(\bar{u}_k) &= f^\pm(\bar{u}_i)_x \Delta x + O(\Delta x^2) \quad k = i-1, i, i+1 \\ d\hat{f}_{i+\frac{1}{2}}^\pm &= \frac{1}{2} f^\pm(\bar{u}_i)_x \Delta x + O(\Delta x^2) \end{aligned} \quad (2.28)$$

Hence in smooth regions away from critical points (critical points are defined here as points for which  $f^+(\bar{u})_x = 0$  or  $f^-(\bar{u})_x = 0$ ), the second and third arguments of the *minmod* functions in (2.17) are asymptotically of the same sign as the first argument and half in

magnitude. Hence the first argument will be picked by the *minmod* function (2.18) for sufficiently small  $\Delta x$ , yielding

$$d\hat{f}_{i+\frac{1}{2}}^{\pm(m)} = d\hat{f}_{i+\frac{1}{2}}^{\pm} \quad (2.29)$$

This guarantees the original high order accuracy of the scheme in such smooth, monotone regions, due to the exponential decay of the off-diagonal entries in  $A^{-1}$  for the type of  $A$  we use [7]. At critical points, the accuracy will degenerate to first order as a generic restriction of all TVD schemes (see, for example, [14]). To overcome this difficulty, we use a modification of the *minmod* function

$$\tilde{m}(a_1, \dots, a_k) = \begin{cases} a_1, & \text{if } |a_1| \leq M\Delta x^2 \\ m(a_1, \dots, a_k), & \text{otherwise} \end{cases} \quad (2.30)$$

where  $M$  is a constant independent of  $\Delta x$ . This modification is discussed in detail in [16] and [4].

With this modification we can obtain schemes which are formally of uniform high order accuracy, equalling the original unlimited scheme, in smooth regions including local extrema. Moreover, we can prove the following

**Proposition 2.5:** The conclusions of Proposition 2.1 and 2.3 are still valid, for any  $n$  and  $\Delta t$  such that  $0 \leq n\Delta t \leq T$ , with TVDM in (2.22) replaced by TVBM (total variation bounded in the means):

$$TV(\bar{u}^n) \leq C \quad (2.31)$$

where  $C$  is independent of  $\Delta t$ , if the *minmod* function  $m$  in (2.17) is replaced by the *modified minmod* function  $\tilde{m}$  defined in (2.30).

□

The choice of the constant  $M$  in (2.30) is related to the second derivative of the solution near smooth extrema. For details, see [16] and [4]. The numerical result is usually not sensitive to the variation of  $M$  in a large range.

### 3 Multi Space Dimensions

For notational simplicity we only consider the two dimensional case (1.1)-(1.2). Three space dimensions do not pose additional conceptual difficulties. As before, we only need to consider the Euler forward time discretization

$$u_{ij}^{n+1} = u_{ij}^n + \Delta t L(u^n)_{ij} \quad (3.1)$$

We again define

$$\bar{u}_{ij} \equiv (A_y A_x u)_{ij} \quad (3.2)$$

so that scheme (3.1) can be left-multiplied by  $A_y A_x$  to become

$$\bar{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{\Delta t}{\Delta x} (A_y B_x f(u^n))_{ij} - \frac{\Delta t}{\Delta y} (A_x B_y f(u^n))_{ij} \quad (3.3)$$



Here and in what follows we will use the commutativity of  $A_x$ ,  $A_y$ ,  $B_x$  and  $B_y$  so that a product can be written in any order. Scheme (3.3) can be written into a conservation form

$$\bar{u}_{ij}^{n+1} = \bar{u}_{ij}^n - \frac{\Delta t}{\Delta x} (\hat{f}_{i+\frac{1}{2},j}^n - \hat{f}_{i-\frac{1}{2},j}^n) - \frac{\Delta t}{\Delta y} (\hat{g}_{i,j+\frac{1}{2}}^n - \hat{g}_{i,j-\frac{1}{2}}^n) \quad (3.4)$$

which involves numerical fluxes  $\hat{f}_{i+\frac{1}{2},j}^n$  and  $\hat{g}_{i,j+\frac{1}{2}}^n$  as local functions of  $u_{kl}^n$ . For example,

$$\begin{aligned} \hat{f}_{i+\frac{1}{2},j} &= \frac{1}{2} A_y (f(u_{i+1,j}) + f(u_{i,j})) \\ \hat{g}_{i,j+\frac{1}{2}} &= \frac{1}{2} A_x (g(u_{i,j+1}) + g(u_{i,j})) \end{aligned} \quad (3.5)$$

for the fourth order central scheme (1.3), etc.. Again, scheme (3.4) resembles a cell-averaged (finite volume) scheme [10]. The  $\bar{u}_{ij}$  defined by (3.2) is a local mean of  $u$ , and a "reconstruction" from  $\bar{u}$  to  $u$

$$u_{ij} = (A_x^{-1} A_y^{-1} \bar{u})_{ij} \quad (3.6)$$

is needed to compute the fluxes  $\hat{f}_{i+\frac{1}{2},j}$  and  $\hat{g}_{i,j+\frac{1}{2}}$  in (3.4).

We remark that the additional costs of implementing scheme (3.4), comparing with the original scheme (3.1), are the two *local* operators  $A_x$  and  $A_y$ . The major part of the cost still consists of the two tridiagonal inversions.

The limiting to obtain nonlinear stability can now be defined in a dimension by dimension fashion: we can use the one-dimensional flux splitting (2.12), for  $f(u)$ , to write the flux  $\hat{f}_{i+\frac{1}{2},j}$  as

$$\hat{f}_{i+\frac{1}{2},j} = \hat{f}_{i+\frac{1}{2},j}^+ + \hat{f}_{i+\frac{1}{2},j}^- \quad (3.7)$$

where  $\hat{f}_{i+\frac{1}{2},j}^\pm$  are again obtained by putting superscripts  $\pm$  in, e.g., (3.5). The remaining definition of the limiting parallels that in Section 2, with a dummy index  $j$  added for the reference  $y$  value: We still start with the differences between the high order numerical fluxes and the first order upwind fluxes

$$d\hat{f}_{i+\frac{1}{2},j}^+ = \hat{f}_{i+\frac{1}{2},j}^+ - f^+(\bar{u}_{ij}); \quad d\hat{f}_{i+\frac{1}{2},j}^- = f^-(\bar{u}_{i+1,j}) - \hat{f}_{i+\frac{1}{2},j}^- \quad (3.8)$$

limit them by

$$\begin{aligned} d\hat{f}_{i+\frac{1}{2},j}^{+(m)} &= m \left( d\hat{f}_{i+\frac{1}{2},j}^+, \Delta_x^+ f^+(\bar{u}_{ij}), \Delta_x^+ f^+(\bar{u}_{i-1,j}) \right) \\ d\hat{f}_{i+\frac{1}{2},j}^{-(m)} &= m \left( d\hat{f}_{i+\frac{1}{2},j}^-, \Delta_x^+ f^-(\bar{u}_{ij}), \Delta_x^+ f^-(\bar{u}_{i+1,j}) \right) \end{aligned} \quad (3.9)$$

where  $\Delta_x^+ v_{ij} \equiv v_{i+1,j} - v_{ij}$  is the forward difference operator in the  $x$  direction, and the *minmod* function  $m$  is defined by (2.18). We then obtain the limited numerical fluxes by

$$\hat{f}_{i+\frac{1}{2},j}^{+(m)} = f^+(\bar{u}_{ij}) + d\hat{f}_{i+\frac{1}{2},j}^{+(m)}; \quad \hat{f}_{i+\frac{1}{2},j}^{-(m)} = f^-(\bar{u}_{i+1,j}) - d\hat{f}_{i+\frac{1}{2},j}^{-(m)} \quad (3.10)$$

and

$$\hat{f}_{i+\frac{1}{2},j}^{(m)} = \hat{f}_{i+\frac{1}{2},j}^{+(m)} + \hat{f}_{i+\frac{1}{2},j}^{-(m)} \quad (3.11)$$

The flux in the  $y$ -direction is defined analogously.

In light of [8] this scheme cannot be TVD in two space dimensions. However we can obtain maximum norm stability through the following

**Proposition 3.1:** Scheme (3.4) with the flux (3.11) satisfies a maximum principle in the means:

$$\max_{i,j} |\bar{u}_{ij}^{n+1}| \leq \max_{i,j} |\bar{u}_{ij}^n| \quad (3.12)$$

under the CFL condition

$$\left[ \max(f^+(u)) + \max(-f^-(u)) \right] \frac{\Delta t}{\Delta x} + \left[ \max(g^+(u)) + \max(-g^-(u)) \right] \frac{\Delta t}{\Delta y} \leq \frac{1}{2} \quad (3.13)$$

where the maximum is taken in  $\min_{i,j} \bar{u}_{ij}^n \leq u \leq \max_{i,j} \bar{u}_{ij}^n$ .

□

In order to obtain maximum norm stability for  $u$ , we need a lemma similar to Lemma 2.2:

**Lemma 3.2:**

If there are two numbers  $0 \leq \delta < 1$  and  $\alpha > 0$ , which are independent of  $N$ , such that the  $N \times N$  matrix  $A = (a_{ij})$  satisfies:

$$\max_{1 \leq i \leq N} \frac{1}{|a_{ii}|} \leq \alpha, \quad \text{and} \quad \sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}| \leq \delta |a_{ii}|, \quad i = 1, \dots, N \quad (3.14)$$

(strongly diagonally dominance for  $A$ ), then the  $L_\infty$  norm of  $A^{-1}$  is bounded independent of  $N$ :

$$\|A^{-1}\|_{L_\infty} \leq \frac{\alpha}{1 - \delta} \quad (3.15)$$

□

For the compact methods we consider, the matrix  $A$  is symmetric. Hence the requirements (2.24) and (3.14) are the same.

We can now use Lemma 3.2 to obtain the maximum norm stability for  $u$ :

**Proposition 3.3:** If a compact scheme (3.4) satisfies the conditions in Proposition 3.1 and Lemma 3.2 for both  $A_x$  and  $A_y$ , then it is stable in the maximum norm. That is,

$$\max_{i,j} |u_{ij}^n| \leq C \quad (3.16)$$

for all  $n \geq 0$  and  $\Delta t > 0$ . Here  $C$  is a constant independent of  $n$  and  $\Delta t$ .

□

This Proposition does not guarantee convergence, but it at least guarantees that the numerical solution will not blow up due to instability.

Under the Assumption 2.4 for both  $A_x$  and  $A_y$ , we can again easily verify that the limiting (3.9) maintains formally the original high order accuracy of the scheme in smooth, monotone regions. The degeneracy of accuracy at critical points can once again be overcome by adopting the *modified minmod* function (2.30) in the limiting (3.9).

## 4 Numerical Examples

To test the behavior of the schemes discussed in Sections 2 and 3, we use the one and two dimensional Burgers equation with smooth initial conditions:

$$\begin{aligned} u_t + \left( \frac{u^2}{2} \right)_x &= 0 \\ u(x, 0) &= 0.3 + 0.7 \sin(x) \end{aligned} \quad (4.1)$$

and

$$\begin{aligned} u_t + \left( \frac{u^2}{2} \right)_x + \left( \frac{u^2}{2} \right)_y &= 0 \\ u(x, y, 0) &= 0.3 + 0.7 \sin(x + y) \end{aligned} \quad (4.2)$$

both with  $2\pi$ -periodic boundary conditions. The solutions will stay smooth initially, then develop shocks which move with time. The exact solution to (4.1) can be obtained by following the characteristics and solving the resulting nonlinear equation using Newton iteration. The exact solution to (4.2) is that of (4.1) with  $x$  replaced by  $x + y$  and  $t$  replaced by  $2t$ . These are standard test problems for scalar nonlinear conservation laws containing shocks. For comparison with finite difference ENO schemes and with finite element discontinuous Galerkin methods, see [17], [4] and [5].

The schemes we test are based on the third order upwind schemes (1.4)-(1.5) coupled with the third order TVD Runge-Kutta time discretization (1.6) (henceforth referred to as the upwind scheme). For the flux splitting (2.12) we use the Lax-Friedrichs splitting (2.14). The time step  $\Delta t^n$  is taken to satisfy a CFL condition

$$\max_i |\bar{u}_i^n| \frac{\Delta t^n}{\Delta x} \leq 0.5 \quad (4.3)$$

in one dimension and

$$\max_{i,j} |\bar{u}_{ij}^n| \left( \frac{\Delta t^n}{\Delta x} + \frac{\Delta t^n}{\Delta y} \right) \leq 0.5 \quad (4.4)$$

in two dimensions. When the *modified minmod* limiter (2.30) is used, the constant  $M$  is taken as 1.

We first test the effect of limiters when the solution is smooth but not monotone. In Figure 1 we plot the  $L_1$  error versus number of grid points, in a log-log scale, at  $t = 0.6$  for the one dimensional case and at  $t = 0.3$  for the two dimensional case. In such scales, the error should be a straight line with slope  $-k$  for a  $k$ -th order method. We can see that the original compact schemes and the schemes with *modified minmod* limiter (2.30) (henceforth referred to as the TVB limiter) give the expected third order accuracy, while the schemes with the *minmod* limiter (2.18) (henceforth referred to as the TVD limiter) give only second order accuracy due to the degeneracy at the critical points.

We then test the effect of limiters when the solution becomes discontinuous. In Figure 2 we show the results of the original compact schemes at  $t = 2$  for the one dimensional case, as well as the result obtained with the TVB limiter (the result obtained with the TVD

limiter is graphically similar to that obtained with the TVB limiter, hence is not shown). We can see over- and under-shoots for the original compact scheme, and monotone shock transition for the result obtained with TVB limiter. In Figures 3, we show the pointwise errors, in a logarithm scale, for the numbers of grid points  $N = 10, 20, 40, 80$  and 160. We can see that the error behaves as expected, with bigger errors for the TVD limiter near the smooth extremum which is close to the shock. The errors for the two dimensional case are similar and are not displayed. In the last picture, Figure 4, we show the surface of the two dimensional solution at  $t = 1$  with  $40 \times 40$  points using the third order upwind method with TVB limiting.

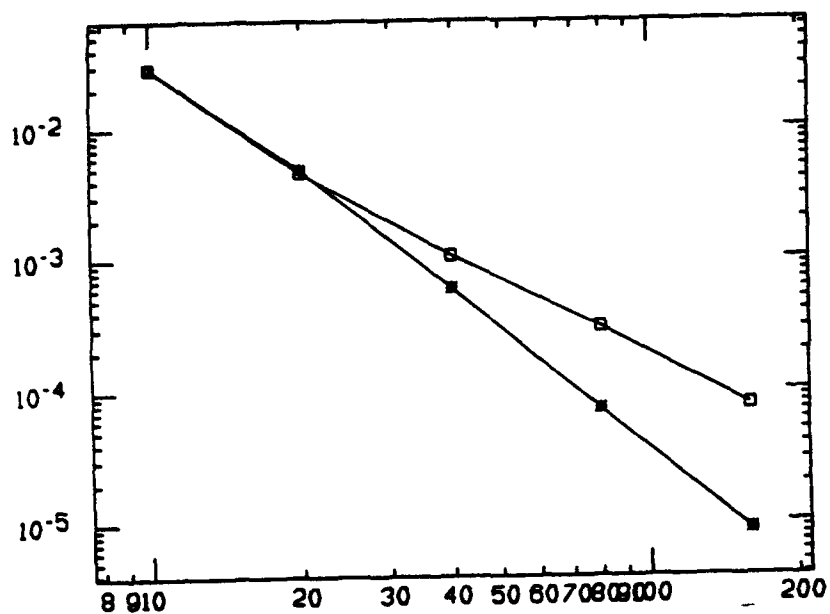
## References

- [1] M. Carpenter, *A high-order compact numerical algorithm for supersonic flows*, In Proceedings of the 12th International Conference on Numerical Methods in Fluid Dynamics, K. W. Morton, ed., Lecture Notes in Physics, v371, Springer-Verlag (1990), pp.254-258.
- [2] M. Carpenter, D. Gottlieb and S. Abarbanel, *The stability of numerical boundary treatments for compact high-order finite-difference schemes*, ICASE Report 91-71 (1991), NASA Langley Research Center.
- [3] M. Ciment and S. Leventhal, *Higher order compact implicit schemes for the wave equation*, Math. Comput., 29 (1975), pp.985-994.
- [4] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: general framework*, Math. Comput., 52 (1989), pp.411-435.
- [5] B. Cockburn, S. Hou and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case*, Math. Comput., 54 (1990), pp.545-581.
- [6] B. Cockburn and C.-W. Shu, *Nonlinearly compact schemes for shock calculations*, SIAM J. Numer. Anal., to appear.
- [7] J. DOUGLAS, JR., T. DUPONT AND L. WHALBIN, *The stability in  $L^q$  of the  $L^2$ -projection into finite element function spaces*, Numer. Math., 23 (1975), pp. 193-197.
- [8] J. Goodman and R. LeVeque, *On the accuracy of stable schemes for 2D scalar conservation laws*, Math. Comput., 45 (1985), pp.15-21.
- [9] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp.357-393.
- [10] A. Harten, *Preliminary results on the extension of ENO schemes to two dimensional problems*, in Proceedings, International Conference on Nonlinear Hyperbolic Problems, Saint-Etienne, 1986. Lecture Notes in Mathematics, C. Carasso et al, eds, Berlin 1987. p.23.

- [11] A. Harten, B. Engquist, S. Osher and S. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes, III*, J. Comput. Phys., 71 (1987), pp.231-303.
- [12] R. Hirsh, *Higher order accurate difference solutions of fluid mechanics problems by a compact differencing technique*, J. Comput. Phys., 19 (1975), pp.90-109.
- [13] S. Lele, *Compact finite difference schemes with spectral-like resolution*, J. Comput. Phys., 103 (1992), pp.16-42.
- [14] S. Osher and S. Chakravarthy, *High resolution schemes and the entropy condition*, SIAM J. Numer. Anal., 21 (1984), pp.955-984.
- [15] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp.1073-1084.
- [16] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Math. Comput., 49 (1987), pp.105-121.
- [17] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock capturing scheme*, J. Comput. Phys., 77 (1988), pp.439-471.
- [18] P. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal., 21 (1984), pp.995-1011.
- [19] R. Vichnevetsky and J. Bowles, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM, Philadelphia, 1982.

Figure 1:  $L_1$  error versus number of grid points in log-log scale for smooth solutions. Stars: compact schemes without limiter; squares: with TVD limiter; diamonds: with TVB limiter.

1(a): Third order upwind scheme, 1D



1(b): Third order upwind scheme, 2D

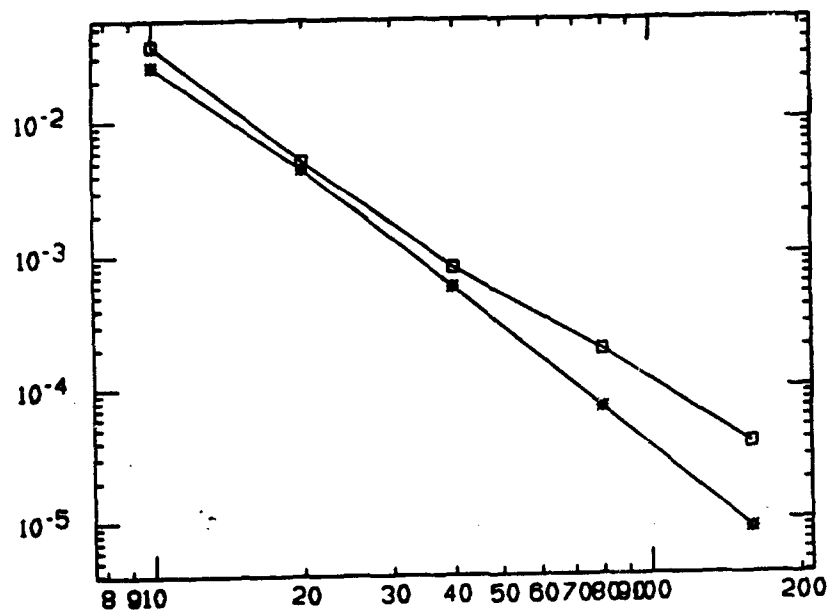
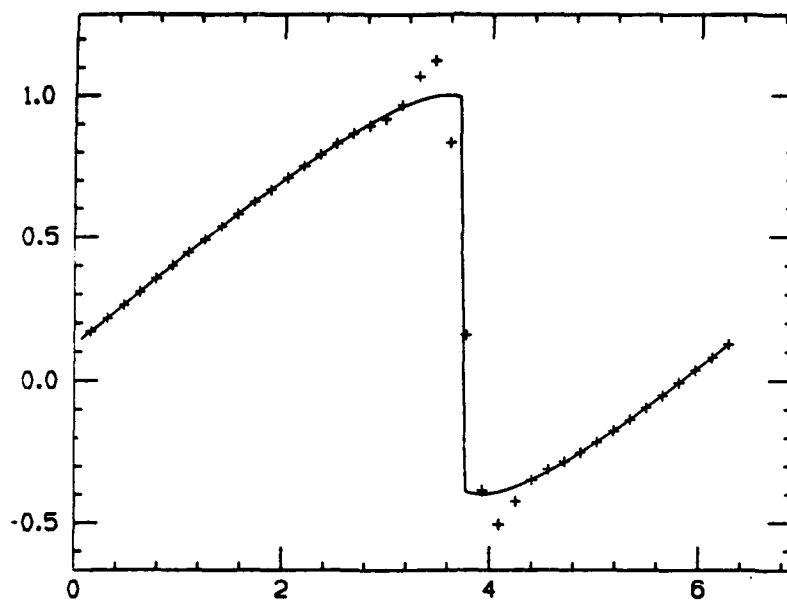


Figure 2: Compact schemes for shocks. Plus signs: computed solution; solid line: exact solution.

2(a): Third order upwind scheme



2(b): Third order upwind scheme with TVB limiter

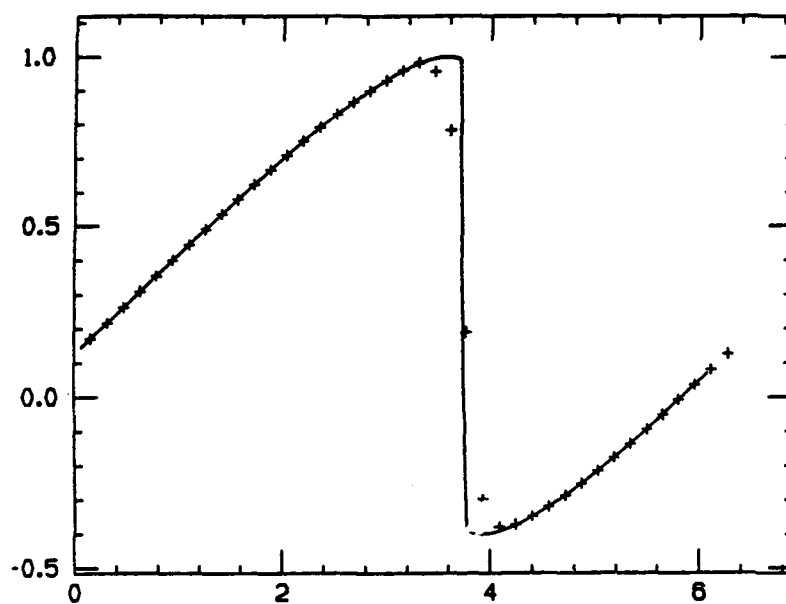
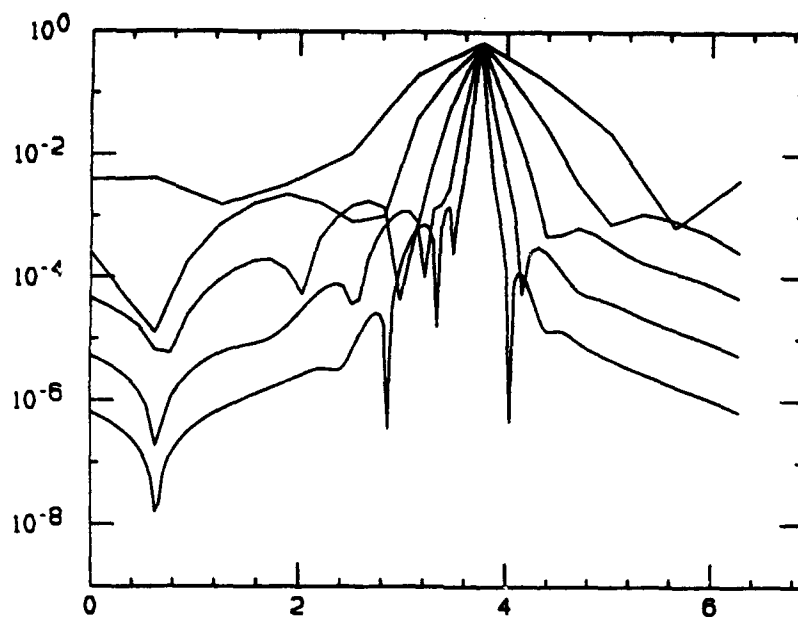
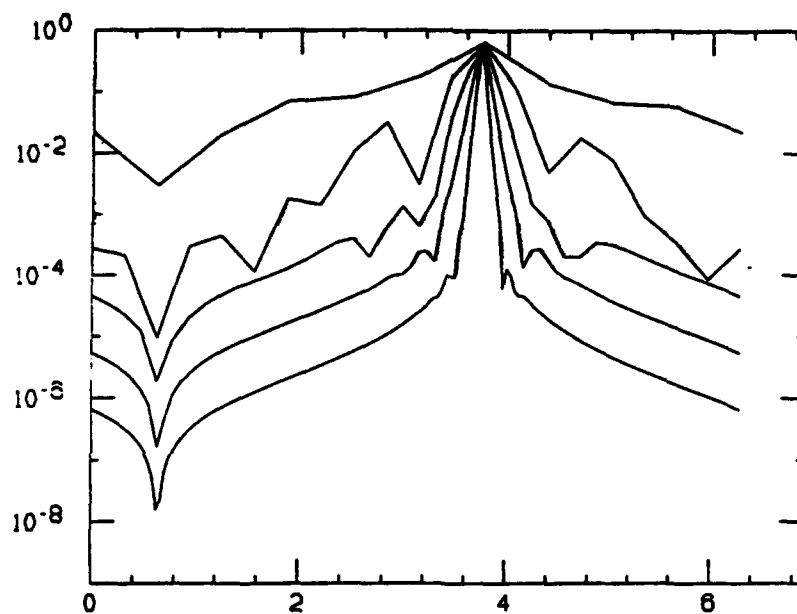


Figure 3: Pointwise error for  $N = 10, 20, 40, 80$  and  $160$  grid points, in a logarithm scale. Third order upwind scheme with limiters for shocks.

3(a): With TVD limiter

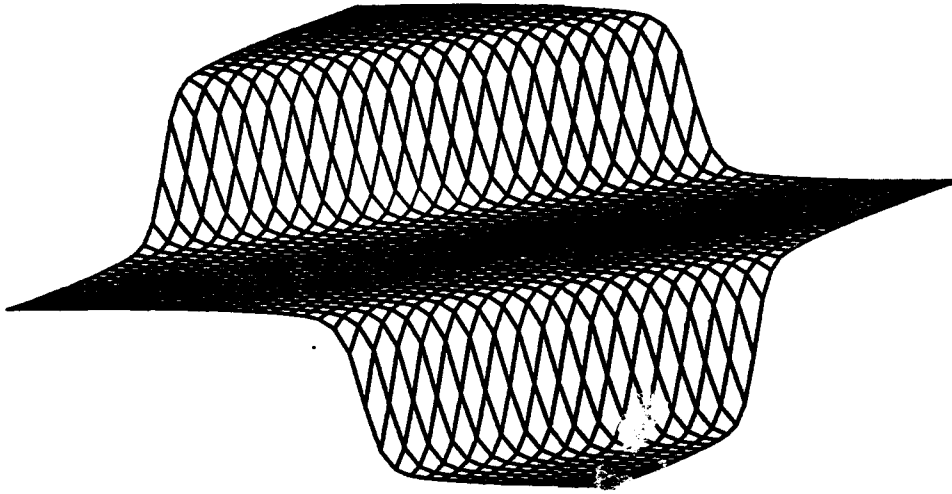


3(b): With TVB limiter





**Figure 4: Surface of third order upwind compact scheme with TVB limiter for shocks.  $40 \times 40$  points.**



# **NUMERICAL WIND TUNNEL TESTING OF PRESSURIZED TENTS\***

Neal E. Blackwell  
Environmental Control and Systems Support Division  
U.S. Army Belvoir Research, Development and Engineering Center  
Fort Belvoir, Virginia 22060-5606

## **ABSTRACT**

Wind tunnel testing of Army temper tents, used in the Chemically Protected Deployable Medical System (CP DEPMEDS), is simulated using a two-dimensional, staggered grid, finite difference approach. This numerical wind tunnel approach is used to predict pressure data to determine lift and drag characteristics of the tents. Wind tunnel blockage due to the tents is less than 10% and wind speeds range from 5 to 70 mph ( $4.6 \times 10^5 \geq Re \leq 6.4 \times 10^6$ ). Length to height ratios for the tents are 6.4 and 9.6, for 19.51 m (64 ft) and 29.26 m (96 ft) length tents, respectively. Values for all the forces are reported for the 19.51 m (64 ft) length tent. Although not reported, the values for the 29.26 m (96 ft) tent length are larger by a factor of 1.5. Drag and lift coefficient values are area specific, therefore, these values are valid for all tent lengths where the two dimensional assumption is valid.

## **INTRODUCTION**

U. S. Army Belvoir RD&E Center (BELVOIR) is continuing to support U. S. Army Natick RD&E Center (NATICK) in the C-100 Air Conditioner program for the Chemically Protected Deployable Medical System (CP DEPMEDS). As a part of that work, BELVOIR is conducting a study of the airflow and drag characteristics of the tents to aid NATICK in the development of an improved tent anchorage system. The magnitude of the forces needed to design the anchorage system are calculated by multiplying the pressure differences across the tent by the surface area. Values for all the forces reported in this study are for 19.51 m (64 ft) length tents. Corresponding values for 29.26 m (96 ft) length tents are larger by a factor of 1.5. Drag and lift coefficients are area specific and hence, are valid for all tent lengths where the two dimensional assumption is valid. This study includes the flow characteristics of a single tent. Future work will include multiple tents positioned downwind. The tents are much longer in length than in width or height and for CP DEPMEDS the tents are positioned parallel to one another and connected at one end by a narrow passageway.

\*Supported by U.S. Army Natick Research Development and Engineering Center.

## **BACKGROUND**

A study of wind loads on the Chemically Protected Deployable Medical System (CP DEPMEDS) tents is needed by U.S. Army Natick Research Development and Engineering Center (NATICK) to assure the correct development of improved tent anchorage systems. Conventional wind tunnel testing is an expensive and time consuming method to obtain wind load data. An alternative and cost-cutting approach is the use of a numerical wind tunnel, where the wind tunnel is simulated on a computer. This inexpensive and time-saving approach is swiftly gaining popularity due to the increasing affordability of computers and improved computational fluid dynamics codes.

In this study, a numerical wind tunnel is used to obtain predictions of velocities and pressures acting on CP DEPMEDS tents (Army temper tents) when wind speeds of 5, 20, 50, and 70 mph strike the upwind side of the tents at a 90 degree angle of attack (broadside).

## **PROCEDURE**

The DEPMEDS tent is modeled as a rigid wall building with a peaked roof. In reality, wind will deform the tent wall into varying shapes. Also, pressures inside and outside of the tent will be transient when the tent materials flap and vibrate due to the wind. Predicting these complex mechanisms is beyond the scope of this study. The following procedures were used to obtain values for engineering calculations leading to the design of a tent anchorage system and are not meant to be a transient simulation of the complex mechanisms involved. The geometrical model of the peaked roof tent is built on a one to one scale, using small, stair step blocks to build the inclined roof. As it turns out, the stair step shape of the roof does not have a major effect on the lift and drag forces on the tent because most the flow over the upwind and downwind roof section is separated from the roof surface. Hence the core flow makes little contact with the roof surface. The height of the vertical tent walls is 1.98 m and the height of the peaked roof is 1.22 m making the total tent height from ground to peak 3.2 m. Wind tunnel height is 55 m and the wind tunnel blockage is 6%, within the recommended value of 10% or less (ASHRAE, 1989). Upwind of the tent is 29.3 m of level ground. Downwind of the tent is 39 m of level ground.

Tent width is assumed a constant value of 6.1 m. This constant value will produce errors at low wind speeds due to the slightly rounded, upwind wall of the pressurized tents. The largest error will be produced by the sharp edge of the upwind wall eaves in the model geometry. However, at high wind speeds, the wind will diminish the roundness of the actual wall because the outside pressure is equal to or exceeds the inside pressure. Roundness of the walls is shown in Figures 1 and 77 of the photograph section of the final report of the Customer Engineering Design Test (Phase II) at Fort Indiantown Gap (Cho and Bryant, 1993). At wind speeds above 16 m/s (36 mph), the upwind wall is expected to form a concave shape that increases with windspeed. The effect of this concavity on drag and lift will be included in future work.

Uniform velocity profiles of magnitudes of 2.24 m/s (5 mph), 8.94 m/s (20 mph), 22.4 m/s (50 mph), and 31.3 m/s (70 mph) are prescribed at the wind tunnel inlet. The flow is assumed to be incompressible because the largest Mach number is 0.09, which is much smaller than  $Ma=0.3$  when compressibility becomes significant. Also, laminar flow simplifications are used to include some viscous effects without the penalty of large computational times associated with turbulent solutions. Turbulent solutions may be used for more detailed, future studies if NATICK desires these results. Turbulent solutions typically double computational time. The inclusion of viscous forces allows the recirculation region, located at the bottom, upwind side of the upwind tent, to be modeled (Figure 1). Viscous forces reduce the approaching wind speed near the ground, produce the recirculation region, and determine the height of the stagnation line. Figure 2 shows the inviscid solution, with the absence of the recirculation region, as compared to the viscous solution in Figure 1. Two dimensional, steady state, conservation equations in Cartesian coordinates are presented below. The conservation of mass is expressed as

$$\frac{\partial(\rho u)}{\partial x} + \frac{\partial(\rho v)}{\partial y} = 0 \quad [1]$$

where terms are defined in the Nomenclature section. The constant density terms are treated as variables due to the general nature of the code CFD2000 1.0, a version of PHOENICS 1.5, which is chosen for this study (Phoenics, 1988). The following equations for the conservation of momentum in the x and y directions are linearized and solved.

$$\frac{\partial(\rho uu)}{\partial x} + \frac{\partial(\rho vu)}{\partial y} = -\frac{\partial p}{\partial x} + \frac{\partial}{\partial x} (\mu_{LAN} \frac{\partial u}{\partial x}) + \frac{\partial}{\partial y} (\mu_{LAN} \frac{\partial u}{\partial y}) \quad [2]$$

$$\frac{\partial(\rho uv)}{\partial x} + \frac{\partial(\rho vv)}{\partial y} = -\frac{\partial p}{\partial y} + \frac{\partial}{\partial x} (\mu_{LAN} \frac{\partial v}{\partial x}) + \frac{\partial}{\partial y} (\mu_{LAN} \frac{\partial v}{\partial y}) \quad [3]$$

To link pressure and flow, CFD2000 uses Semi-Implicit-Method for Pressure-Linked-Equations (SIMPLE), developed by Patankar (1980), in conjunction with the staggered grid method. The staggered grid method is used to prevent oscillations of the pressure field. As a result, pressures are calculated at cell nodes and velocities are calculated at cell faces.

Horizontal grid resolution is 1.63 m/cell in region 1, 0.24 m/cell in region 2, and 1.63 m/cell in region 3 (Figure 3). Vertical grid resolution is 0.25 m/cell in region A, 0.17 m/cell in region B, and 2.5 m/cell in region C (Figure 3). The discretized forms of equations 1, 2, and 3 are solved at 4,234 cell nodes, totalling 12,702 simultaneous equations.

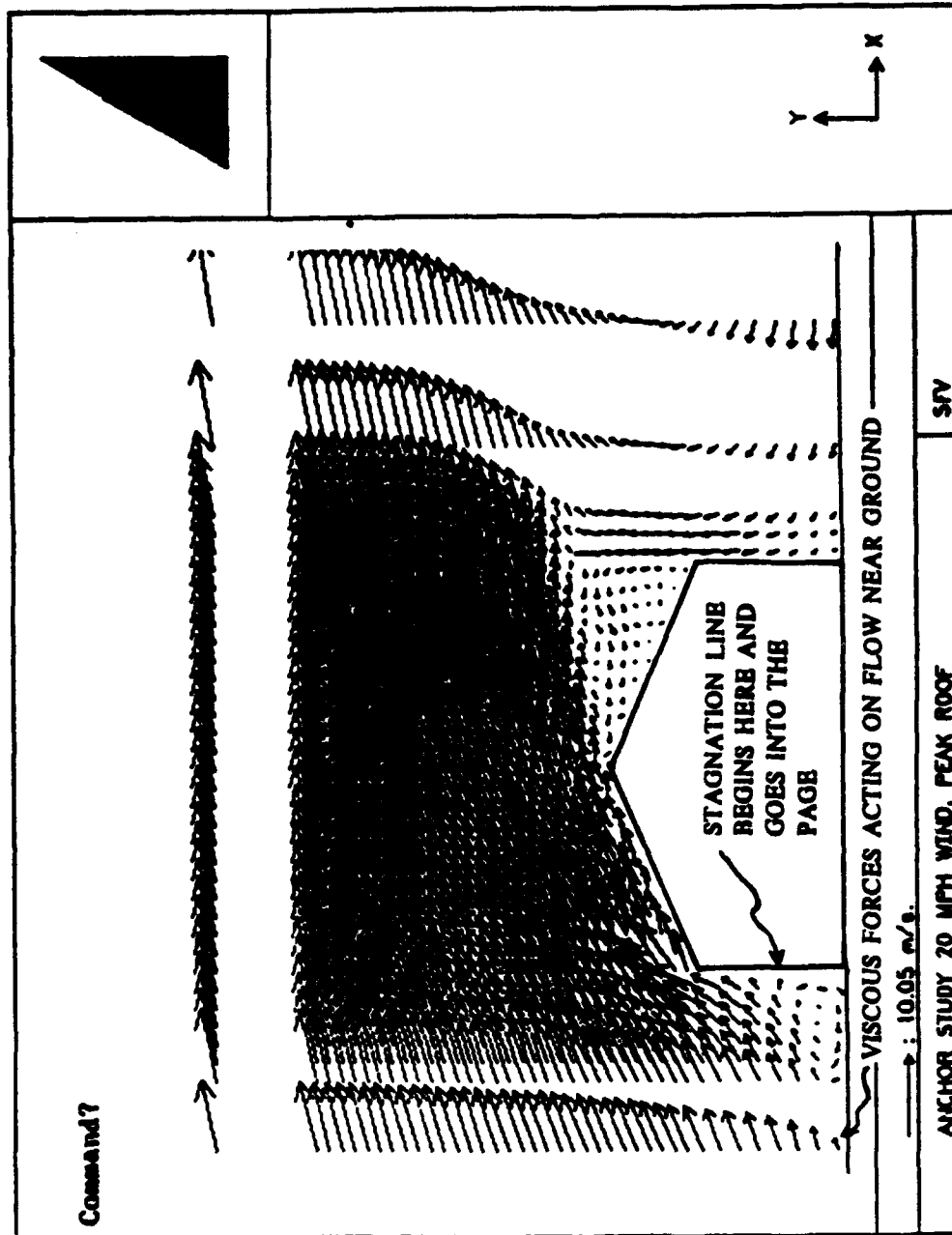


FIGURE 1. VELOCITY PROFILE, STAGNATION LINE AND VISCOUS EFFECTS  
ON ARMY TEMPER TENT.

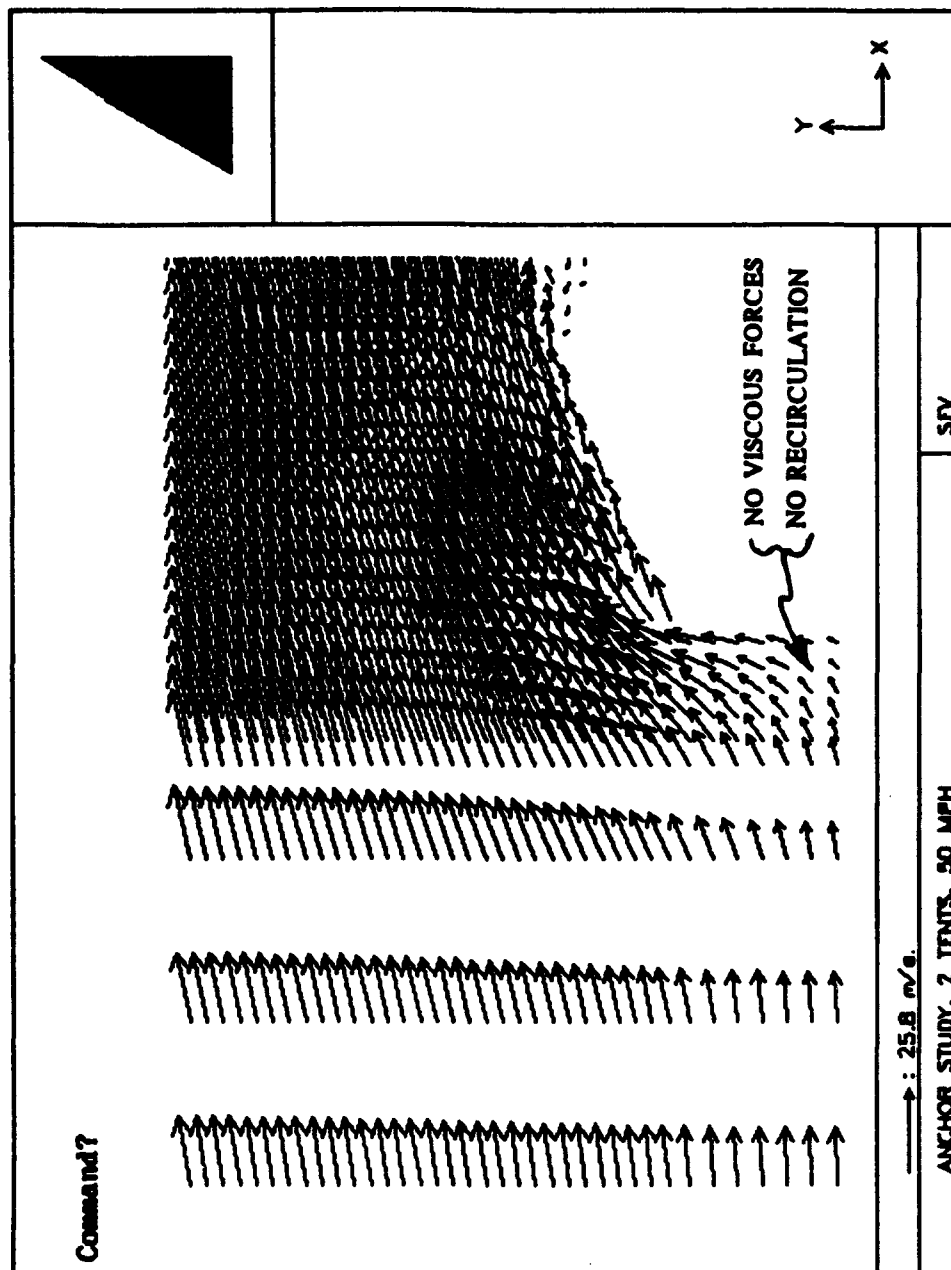


FIGURE 2. INVISCID SOLUTION RESULTING FROM EULER'S EQUATIONS.

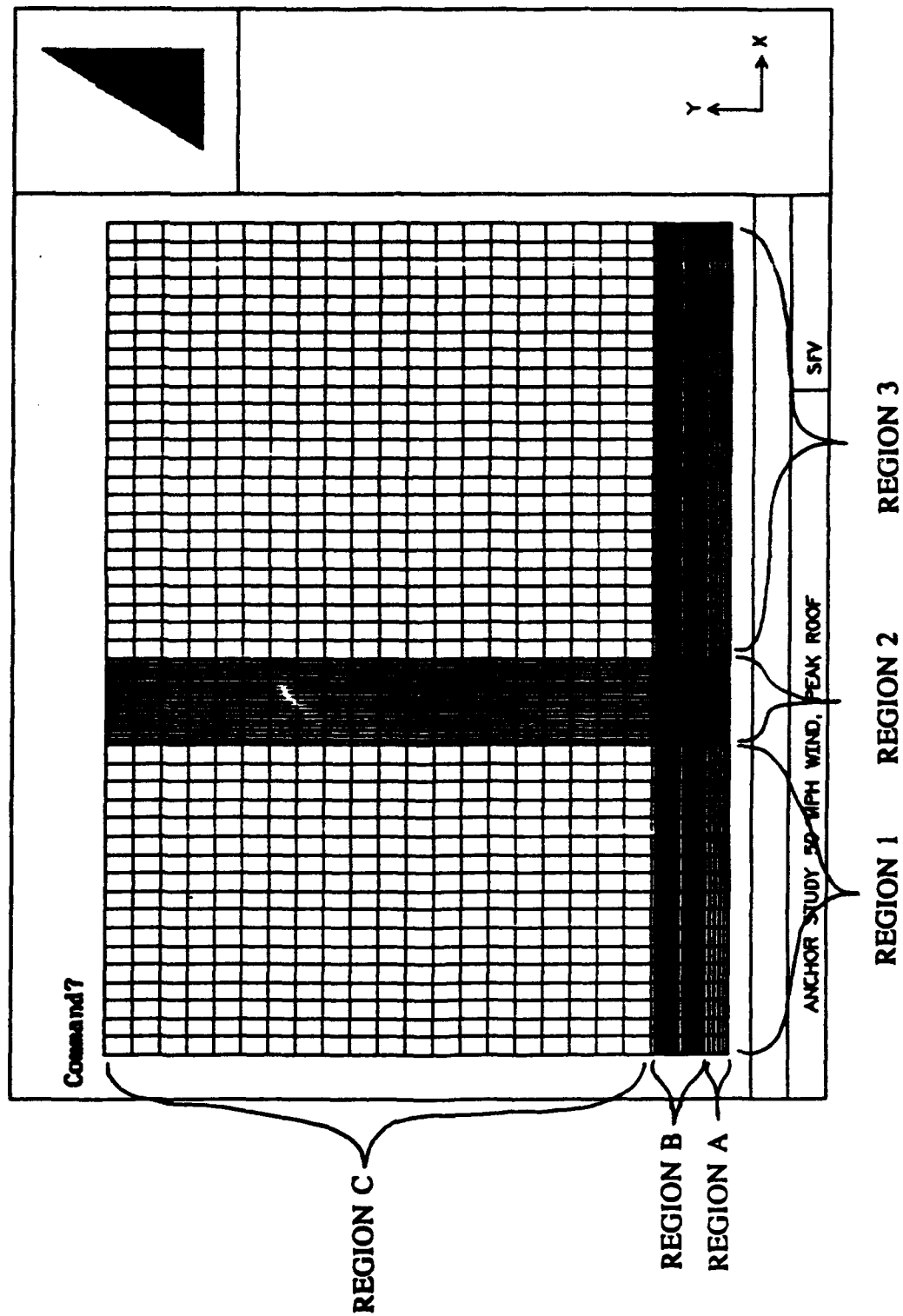


FIGURE 3. REGIONS OF DISCRETIZATION.

## **RESULTS**

Solutions converge within 2500 to 3000 iterations and graphical examples of the results are in Figures 4 and 5. Approximately three hours of computer run time is required for convergence using a 486 personal computer operating at 33 MHZ. The 32 bit Lahey F77L-EM/32 FORTRAN Compiler version 4.02 with the Lahey Ergo OS/386 DOS Extender is used.

Numerical wind tunnel results are interpreted using graphs that show velocity vectors and pressure contours. The direction of the velocity vectors is indicated by the angle of the arrows and the magnitude of the vector, in m/s, is indicated by the length of the arrow. Pressure magnitudes, in pascals, are indicated as varying shades of black where the pressure increases as the shade lightens. Values assigned to the shaded contours are printed in the legend located on the right side of the pressure contour plot. An example of how to interpret the pressure contour predictions is to note the light shade of the contour on the lower upwind side of the tent and then note the darker shade contour on the lower downwind side. Noting the pressure shaded legend on the right side of the page, the pressure values may be found; and the difference between the two values is the maximum pressure drop across the tent. The individual pressures are relative, not absolute, due to the  $\partial P/\partial x$  and  $\partial P/\partial y$  terms in the x and y momentum equations. Hence, the difference between two relative pressures yields the correct pressure difference.

Figures 4 and 5 are representative of the flow patterns and pressure contours over the computational domain. Three main flow regimes exist and these include separated flow, recirculation, and reattachment (Figures 4, 5, 6 and 7). Flow separates near the peak of the roof and recirculates on the downwind side of the roof and tent wall. The separated flow reattaches to the ground downwind, and these distances and ratios (reattachment length/tent height)(L/H) are given in Table 1. Objects located in the separation region are shielded from wind loads and are less likely to be blown around or damaged during high wind gusts. Height of the separation region decreases downwind; therefore, the tallest objects to be shielded should be positioned closest to the downwind wall.

### **Drag and Lift**

Drag and lift forces and the corresponding coefficients are given in Table 1 and are important for tent anchorage. The drag and lift coefficients versus wind speed and Reynolds Number are a constant value (Table 1) and are defined as

$$C_D = \frac{Drag}{1/2\rho V^2 A} \quad [4]$$

$$C_L = \frac{Lift}{1/2\rho V^2 A} \quad [5]$$



a

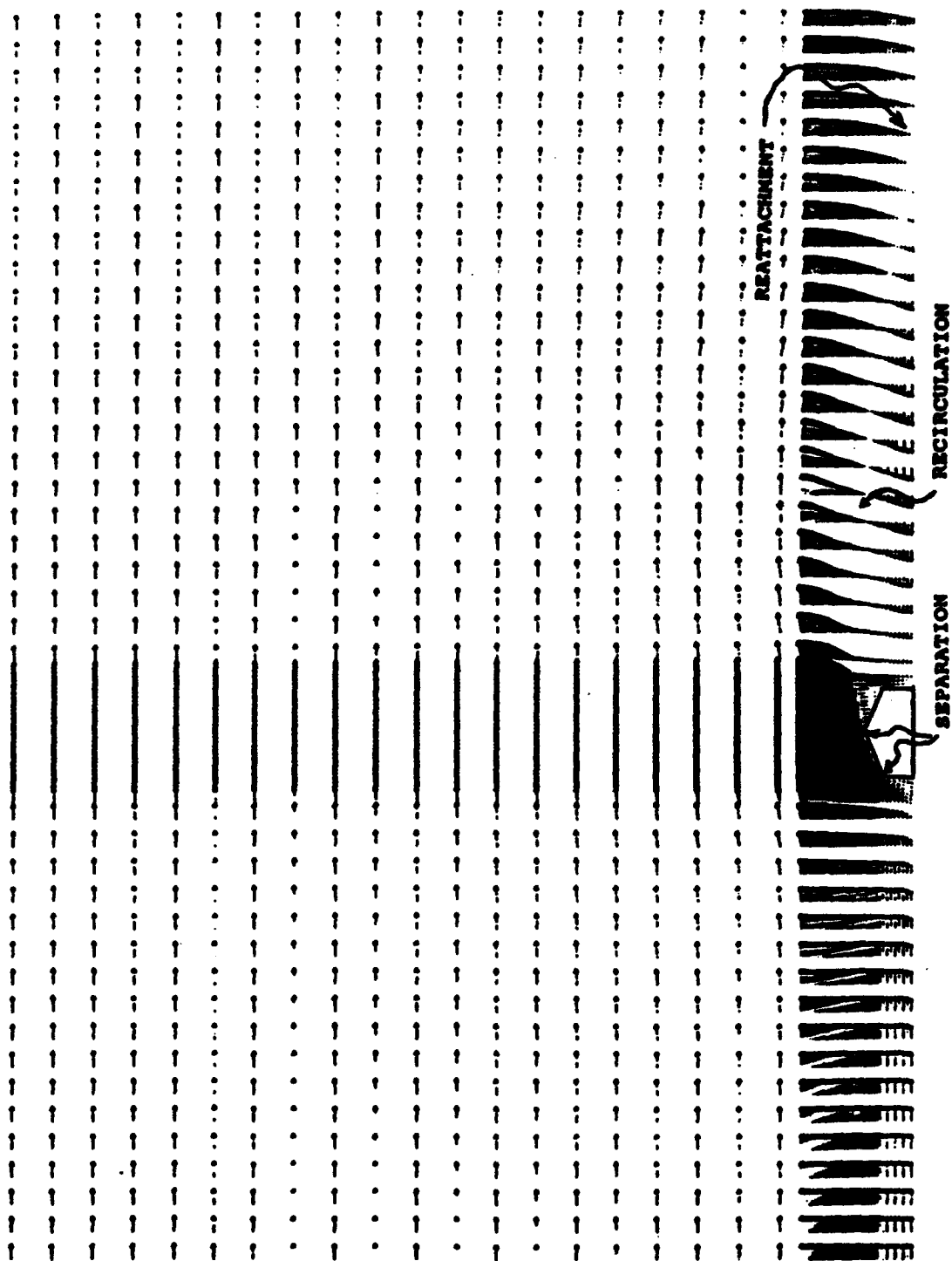


FIGURE 4. VELOCITY VECTORS RESULTING FROM 50 MPH WIND.

a

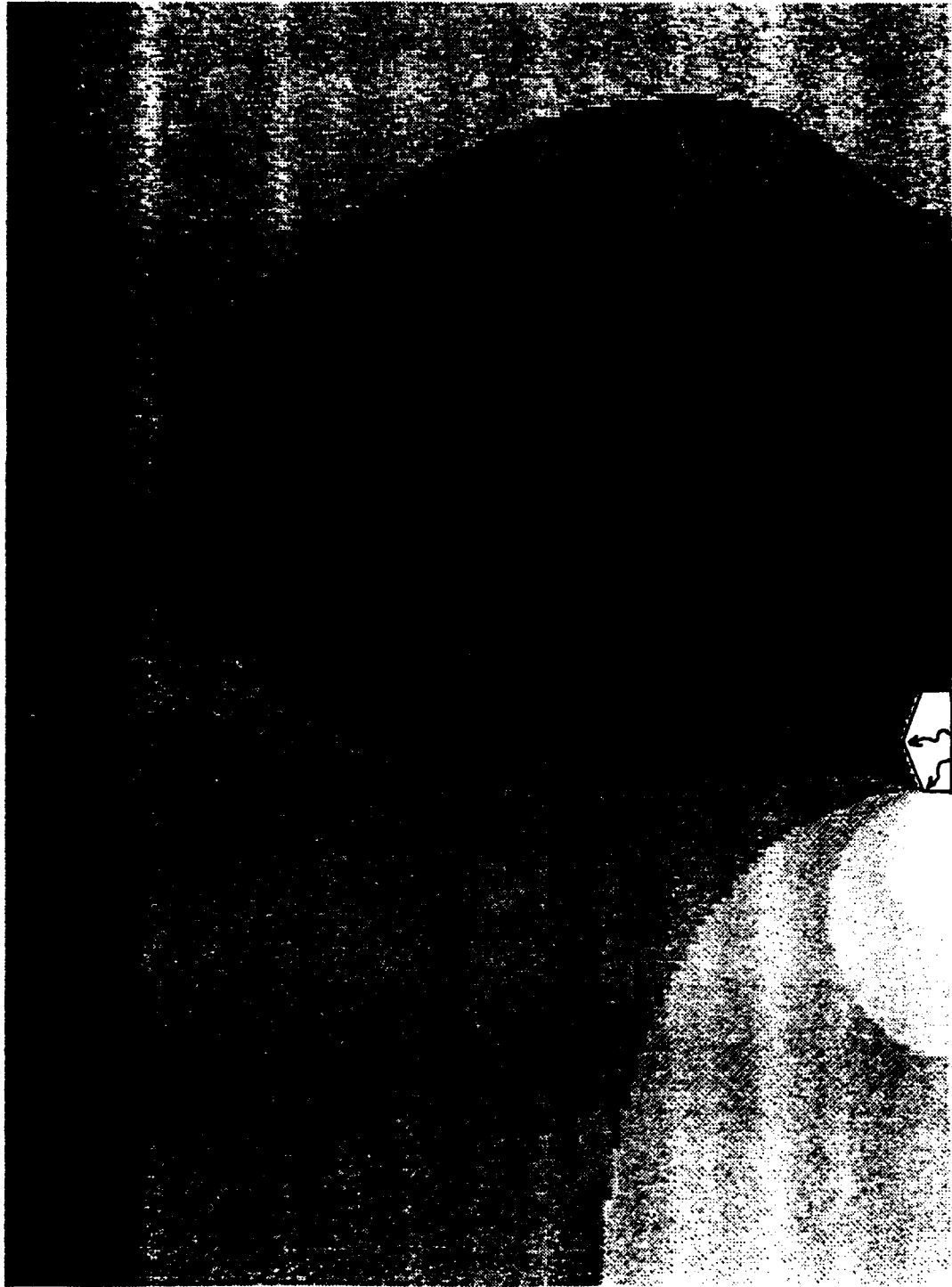


FIGURE 5. PRESSURE CONTOURS RESULTING FROM 50 MPH WIND.

REATTACHMENT

RECIRCULATION

SEPARATION

SFV

-278  
-237  
-196  
-155  
-114  
-73  
-31  
10  
51  
92  
133  
174  
215  
256  
297

Y

X

a

-278  
-237  
-196  
-155  
-114  
-73  
-31  
10  
51  
92  
133  
174  
215  
256  
297

Y

X



SFV

FIGURE 6. COMPOSITE PLOT OF VELOCITY AND PRESSURE.

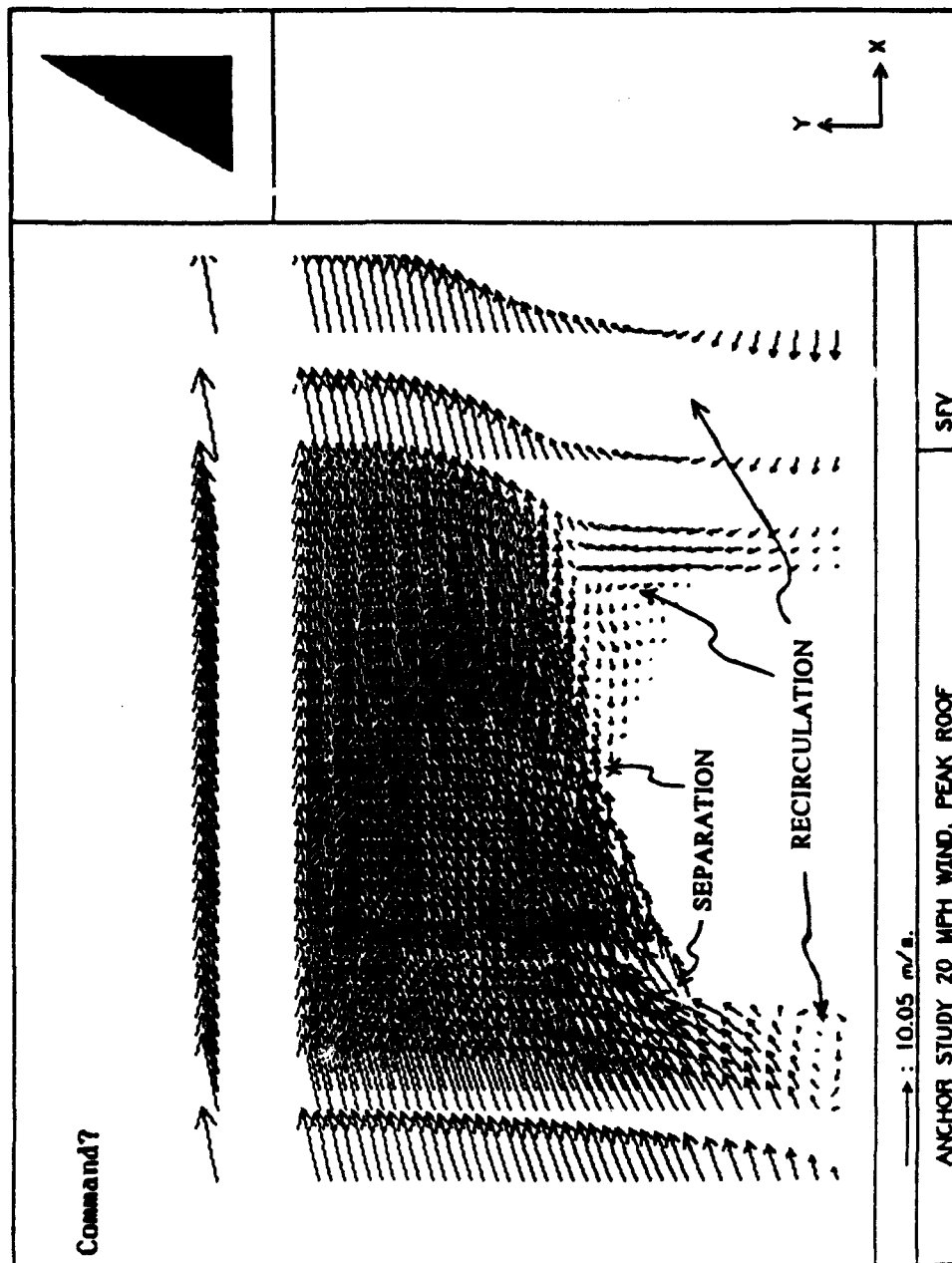


FIGURE 7. MAGNIFIED VIEW OF SEPARATION AND RECIRCULATION REGIONS.

Hence, drag and lift may be calculated for any wind speed between 5 and 70 mph. As equations 4 and 5 and Figure 8 show, drag and lift increase quadratically with respect to wind speed. Drag is calculated by summing the products of the pressure differences, across the tent width by the cell areas of the upwind wall and the vertical projection of the roof. Drag is the horizontal force that the tent anchorage system must support.

Table 1. Drag, Lift and Reattachment Parameters as a function of Wind Speed and Reynolds Number (based on tent height).

Wind Speed km/h (mph)	8.0 (5)	32.2 (20)	80.5 (50)	112.6 (70)
Reynolds No.	$4.56 \times 10^5$	$1.82 \times 10^6$	$4.56 \times 10^6$	$6.39 \times 10^6$
Drag Coeff	0.78	0.77	0.77	0.77
Lift Coeff	0.45	0.45	0.44	0.44
Drag N (lbf)	140 (30)	2260 (510)	14140 (3180)	27580 (6200)
Lift N (lbf)	160 (35)	2500 (560)	15470 (3480)	30110 (6770)
Reattachment Distance m (ft)	36.2 (119)	36.5 (120)	37.1 (122)	37.4 (123)
Ratio L/H	11.3	11.4	11.6	11.7

Lift is calculated by summing the products of the pressure differences across the height of the tent by the cell areas of the horizontal projection of the roof. The pressure at the bottom of the tent is assumed to be at atmospheric pressure. Flow separation at the leading edge of the roof (upwind eaves) produces a low pressure region that tends to lift the upwind portion of the tent (Figure 7). However, the prominent flow separation region exists over the downwind roof section, and it produces a low pressure region that tends to lift the downwind portion of the tent. Lift is the vertical force that the anchorage system must support.

#### Forces Tending to Cause Roof and Wall Blowout

During the Customer Engineering Design Test (Phase II) that was conducted during the summer of 1992, the U.S Army Combat Systems Test Activity (CSTA) reported several occurrences of wall blowout during pressurization (Cho, 1992). Wall blowouts were prevented by insuring that the overpressure value did not exceed 0.6 inches of water (IW). In a no wind condition, the CP DEPMEDS recommended overpressure of 0.6 IW produces a 5770 N (1300 lbf) force acting perpendicular to each 1.98m x 19.51m upwind and downwind wall. The pressure difference across all the walls and the roof are equal. However, under wind load conditions, the pressure difference across the tent walls and roof varies

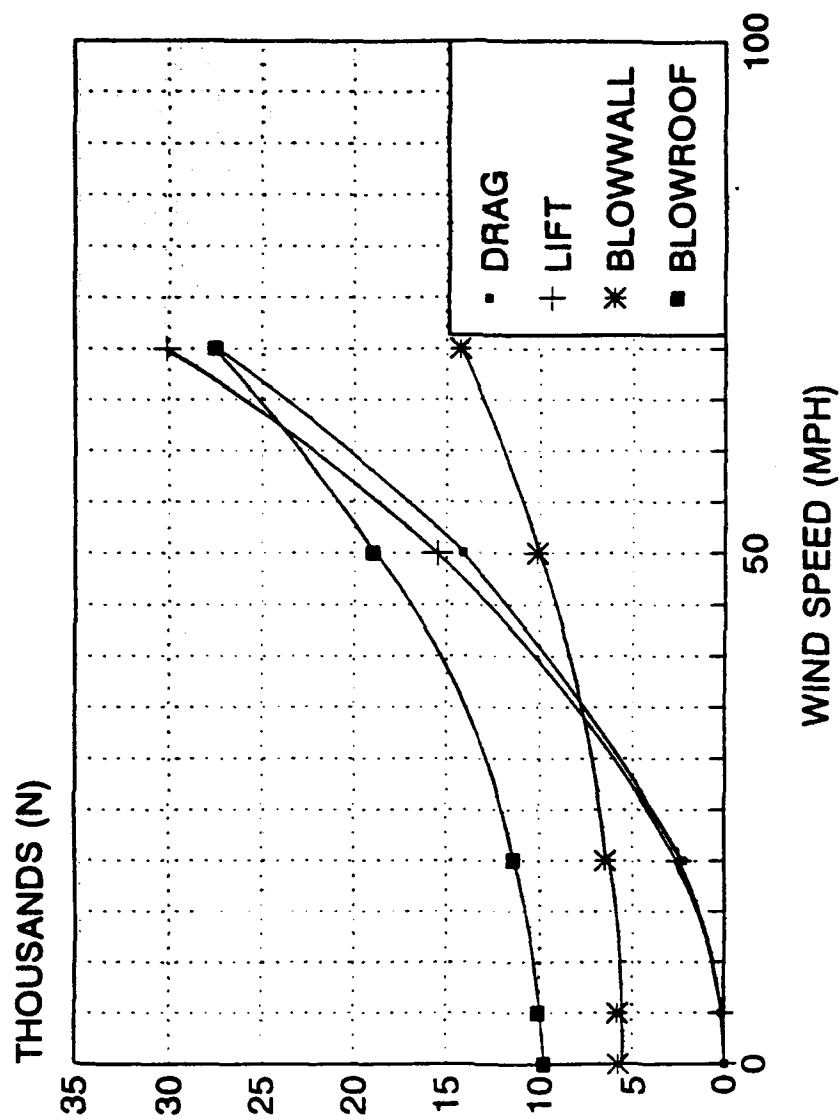


FIGURE 8. WIND LOADS ON ARMY TEMPER TENT.

around the tent, with the minimum value existing on the upwind wall and the maximum value existing on the downwind wall. The maximum pressure difference occurs on the downwind side because a relatively high pressure exists in the tent and a relatively low pressure exists outside the tent. The difference between the two pressures is relatively large. Resultant forces on the downwind wall due to tent pressurization and wind loads are given in Table 2 and Figure 8. The downwind wall is the section most vulnerable to blowout.

The downwind roof section is the next most vulnerable section to blowout. In a no wind condition, the overpressure produces a 9575 N (2150 lbf) force acting perpendicular to each roof section. Under wind load conditions, the projection of the lift forces perpendicular to the roof plus the overpressure force of 9575 N combine to form a larger force that tends to blowout the downwind roof section (Table 1 and Figure 8). Resultant forces on the downwind roof section are given in Table 2 and Figure 8.

Table 2. Forces Normal to the Downwind Roof and Wall Sections.

Wind Speed (mph)	Force Normal to Downwind Roof Section N (lbf)	Force Normal to Downwind Wall N (lbf)
0	9575 (2150)	5770 (1300)
5	10100 (2270)	5830 (1310)
20	11460 (2580)	6490 (1460)
50	19024 (4280)	10150 (2280)
70	27550 (6190)	14280 (3210)

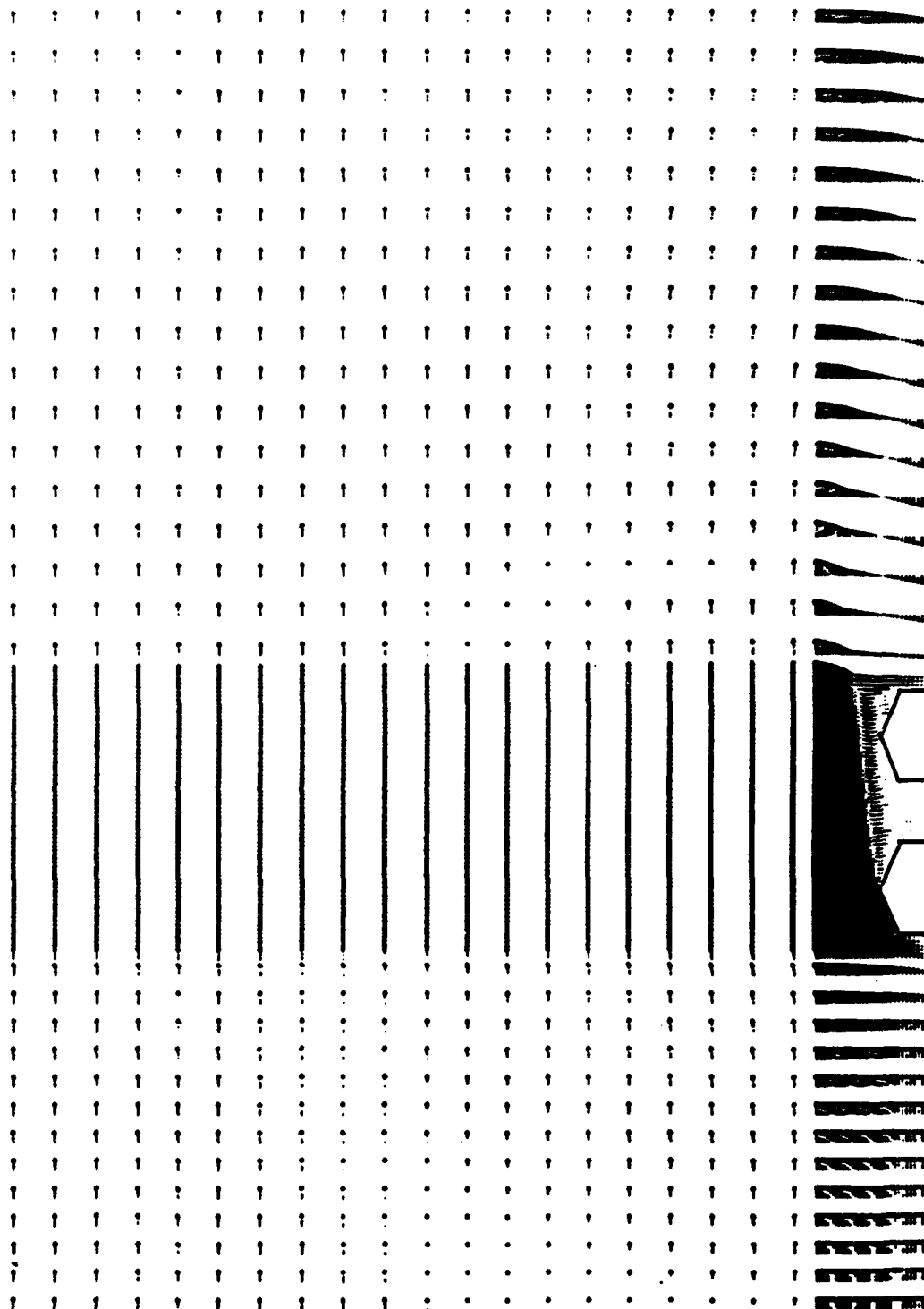
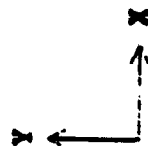
### CONCLUSIONS

Drag and lift forces are predicted for 5, 20, 50 and 70 mph wind speeds. The constant values for the drag and lift coefficients allow drag and lift to be calculated for any wind speed between 5 and 70 mph and for any length tent where the two dimensional assumption is valid. Also, Figure 8 may be used to predict the forces tending to blow out the downwind roof and wall sections. These forces will aid in the development of stronger seams for the NBC tent liner. Forces are recorded to provide data for development of improved tent anchorage systems.

### FUTURE WORK

BELVOIR is presently predicting air flow and drag characteristics of multiple CP DEPMEDS tents positioned downwind (Figures 9 and 10). Introduction of a second tent doubles computational time, hence, Euler equations for inviscid flow are used to reduce computational time.

a



SFV

FIGURE 9. VELOCITY VECTORS OF TWO TENTS IN PARALLEL.



a

-236  
-195  
-155  
-114  
-74  
-33  
7  
47  
88  
128  
169  
209  
250  
290  
330

y

x



FIGURE 10. PRESSURE CONTOURS OF TWO TENTS IN PARALLEL.

SFV

Preliminary results indicate that the outer two tents of the CP DEPMEDS complex block the air from the inner tents. The requirement for expensive, heavy duty, anchorage equipment may be limited to the outer tents. The inner tents can be fitted with less expensive, lighter, anchorage equipment. This helps to reduce cost and weight of the anchorage system.

Also, the slower air flowing between tents may function like a pneumatic separator in sandy, desert regions where dust and sand storms are common. The particles being transported in the separated regime could settle between tents as they contact the slower recirculation regimes between tents (Figure 9). This could produce sand, dust, or snow drifts.

Finally, we expect the high pressure that exists on the upwind wall of the upwind tent and the low pressure that exists between tents to combine and defeat the overpressure system for wind speeds above 16 m/s (36 mph). This will cause contaminants to enter the upwind tent and quickly flow, by convective transport, into each of the downwind tents. This will happen because the downwind tents are engulfed in the low pressure region shown in the shaded area in Figure 10. Convective transport is expected to spread the contaminants from the upwind tent to the downwind tents so quickly that medical personnel will not have time to isolate the downwind tents.

### **ACKNOWLEDGEMENTS**

The author thanks U.S. Army Natick Research, Development and Engineering Center for support of this work. Special thanks to Andra Kirsteins of NATICK and John Waddick of BELVOIR for their leadership in this study.

### **REFERENCES**

- Phoenics of North America. 1988. Perpetual License Agreement. Atlanta, GA.
- ASHRAE. 1989. ASHRAE HANDBOOK - FUNDAMENTALS, American Society of Heating, Refrigeration and Air-Conditioning Engineers, Inc., Atlanta, GA., p 14.16.
- Cho N.K. and R.P. Bryant. 1993. Customer Engineering Design Test (Phase II) of the Chemically Protected Deployable Medical System (CP DEPMEDS). U.S. Army Combat Systems Test Activity, Aberdeen Proving Ground, MD.
- Cho N.K. 1992. Personal Communication. U.S. Army Combat Systems Test Activity, Aberdeen Proving Ground, MD.
- Patankar, S. V., 1980, Numerical Heat Transfer and Fluid Flow, Hemisphere Publishing Corporation, New York.

## **NOMENCLATURE**

$\rho$  = air density

$V$  = air velocity

$x, y, z$  = cartesian coordinates

$u, v, w$  = velocity components

$\partial$  = differential operator

$P$  = pressure

$\mu_{LAM}$  = laminar viscosity

$A$  = projected area

# THREE-DIMENSIONAL FINITE ELEMENTS WITH MULTIPLE-QUADRATURE-POINTS

Wing Kam Liu, Yu-Kan Hu and Ted Belytschko

Department of Mechanical Engineering  
Northwestern University  
Evanston, IL 60208

**ABSTRACT.** New multiple-quadrature-point underintegrated finite elements with hourglass control are developed. The elements are underintegrated to avoid volumetric and shear locking and save computational time. An approach for hourglass control is proposed such that the stabilization operators are obtained simply by taking the partial derivatives of the generalized strain rate vector with respect to the natural coordinates so that the elements require no stabilization parameter. To improve accuracy over the traditional one-point-quadrature elements, several quadrature points are used to integrate the internal forces, especially for tracing the plastic fronts in the mesh during loading and unloading in elastic-plastic analysis. Four-point quadrature are proposed for use in the two and three dimensional elements. Other multiple quadrature points can also be employed. Several numerical examples such as thin beam, plate and shell problems are presented to demonstrate the applicability of the proposed elements.

**INTRODUCTION.** In large scale finite element analyses, thousands of elements and large computer memory demands are required to obtain the detailed information for engineering design or process control. In these analyses, computational costs are mostly determined by the efficiency of the elements, especially for nonlinear problems. Perhaps, the most efficient elements are the one-point-quadrature elements with hourglass control developed by Flanagan and Belytschko [1], Belytschko [2] and Belytschko *et al.* [3]. The mesh instability associated with the under-integrated elements is controlled by adding a stabilization to the one-point quadrature element. The stabilization terms are obtained by ensuring the consistency of the finite element equations and its magnitude is controlled by a user controlled stabilization parameter. Liu and Belytschko [4] also develop a one-point-quadrature element for heat conduction problems. In this work, the stabilization parameter is determined by solving an eigenvalue problem. The relationships between the stabilization parameter and finite difference formulae and full integration finite elements are discussed in the same paper. In Belytschko *et al.* [5], the Hu-Washizu variational principle is used to examine the magnitude of the stabilization parameters. More recently, an assumed strain stabilization of the four-node quadrilateral element and the eight-node hexahedral element with one-point-quadrature, where the stabilization parameters are not required, was proposed by Belytschko and Bindeman [6,7].

An alternative approach for hourglass control is proposed by Liu *et al.* [8], in which the resulting stabilization matrix requires no stabilization parameter. It is shown that the stabilization vector  $\gamma$  can be obtained simply by taking the partial derivatives of the generalized strain vector with respect to the natural coordinates. The strain vector is therefore approximated by the combination of a constant part and other parts involving strain derivatives. However, shear-related locking phenomena are not taken into consideration and no three dimensional result is reported in their study. Another technique is the so-called directional reduced integration proposed by Koh and Kikuchi [9] based on the procedure given in Liu *et al.* [8]. In contrast to selective reduced integration, where certain parts of internal virtual work are underintegrated uniformly in all directions, the directional reduced integration underintegrates in certain directions. Numerical

---

\* This research is sponsored by an ARO Grant number DAAL03-91-G-0016 and National Center for Supercomputing Applications at Urbana-Champaign.

examples show that this technique is effective for two dimensional problems. The authors also extend the same technique to develop three dimensional plate/shell elements; however, hourglass modes are found in those elements.

In this paper, two and three dimensional underintegrated elements based on procedures similar to that proposed by Liu *et al.* [8] are developed. The emphasis is placed on the avoiding of locking and the removal of spurious singular modes. These elements are applicable to beam, plate and shell bending problems, and more importantly they are suitable to analyze metal forming processes.

**EIGHT-NODE HEXAHEDRAL ELEMENT.** Let us consider an eight-node hexahedral element as shown in Figure 1. The spatial coordinates  $x_i$  and the velocity components  $v_i$  in the element are approximated by linear combinations of nodal values  $x_{ia}$  and  $v_{ia}$ , and shape functions  $N_a(\xi, \eta, \zeta)$  as follows:

$$x_i = \sum_{a=1}^{NEN} N_a(\xi, \eta, \zeta) x_{ia} \quad NEN=8 \quad (1)$$

$$v_i = \sum_{a=1}^{NEN} N_a(\xi, \eta, \zeta) v_{ia} \quad (2)$$

$$N_a(\xi, \eta, \zeta) = \frac{1}{8}(1+\xi_a\xi)(1+\eta_a\eta)(1+\zeta_a\zeta) \quad (3)$$

where the subscripts "i" and "a" denote coordinate component ranging from one to three and the element node number, ranging from one to eight, respectively. The referential coordinates  $\xi$ ,  $\eta$  and  $\zeta$ , of node a are denoted by  $\xi_a$ ,  $\eta_a$  and  $\zeta_a$ , respectively.

For the purpose of identifying the deformation modes of the element, let us define the gradient submatrices  $B_a(0)$  and other column vectors as:

$$B_a(0) = \begin{bmatrix} N_{a,x}(0) \\ N_{a,y}(0) \\ N_{a,z}(0) \end{bmatrix} = \begin{bmatrix} b_{1a} \\ b_{2a} \\ b_{3a} \end{bmatrix} \quad (4)$$

$$\underline{s}^t = [1, 1, 1, 1, 1, 1, 1, 1] \quad (5)$$

$$\underline{x}^t = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8] \quad (6)$$

$$\underline{y}^t = [y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8] \quad (7)$$

$$\underline{z}^t = [z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8] \quad (8)$$

$$\underline{h}_1^t = [1, -1, 1, -1, 1, -1, 1, -1] \quad (9)$$

$$\underline{h}_2^t = [1, -1, -1, 1, -1, 1, 1, -1] \quad (10)$$

$$\underline{h}_3 = [1, 1, -1, -1, -1, -1, 1, 1] \quad (11)$$

$$\underline{h}_4 = [-1, 1, -1, 1, 1, -1, 1, -1] \quad (12)$$

$$\underline{\xi}^t = [-1, 1, 1, -1, -1, 1, 1, -1] \quad (13)$$

$$\underline{\eta}^t = [-1, -1, 1, 1, -1, -1, 1, 1] \quad (14)$$

$$\underline{\zeta}^t = [-1, -1, -1, -1, 1, 1, 1, 1] \quad (15)$$

where  $\underline{x}$ ,  $\underline{y}$  and  $\underline{z}$  are the nodal coordinates and  $\underline{h}_1$  is the  $\xi\eta$ -hourglass vector,  $\underline{h}_2$  the  $\xi\zeta$ -hourglass vector,  $\underline{h}_3$  the  $\eta\zeta$ -hourglass vector and  $\underline{h}_4$  the  $\xi\eta\zeta$ -hourglass vector.

The Jacobian matrix evaluated at the center of an element can be shown to be:

$$I(\mathcal{Q}) = [J_{ij}] = \frac{1}{8} \begin{bmatrix} \underline{\xi}^t \underline{x} & \underline{\xi}^t \underline{y} & \underline{\xi}^t \underline{z} \\ \underline{\eta}^t \underline{x} & \underline{\eta}^t \underline{y} & \underline{\eta}^t \underline{z} \\ \underline{\zeta}^t \underline{x} & \underline{\zeta}^t \underline{y} & \underline{\zeta}^t \underline{z} \end{bmatrix} \quad i, j = 1, 2, 3 \quad (16)$$

The determinant of the Jacobian matrix is denoted by  $J_0$  and the inverse matrix of  $I(\mathcal{Q})$  is given by  $\mathcal{Q}$ :

$$\mathcal{Q} = [\mathcal{Q}_{ij}] = I^{-1}(\mathcal{Q}) \quad (17)$$

The gradient vectors  $\underline{h}_1$ ,  $\underline{h}_2$  and  $\underline{h}_3$  (which are evaluated at  $(\mathcal{Q})$  in equation (4), can be shown to be

$$\underline{h}_1 = \{b_{1a}\} = \frac{1}{8} [D_{11}\underline{\xi} + D_{12}\underline{\eta} + D_{13}\underline{\zeta}] \quad (18)$$

$$\underline{h}_2 = \{b_{2a}\} = \frac{1}{8} [D_{21}\underline{\xi} + D_{22}\underline{\eta} + D_{23}\underline{\zeta}] \quad (19)$$

$$\underline{h}_3 = \{b_{3a}\} = \frac{1}{8} [D_{31}\underline{\xi} + D_{32}\underline{\eta} + D_{33}\underline{\zeta}] \quad (20)$$

The strain rate  $\dot{\underline{\epsilon}}$  is approximated by expanding it in a Taylor series about the element center up to bilinear terms:

$$\begin{aligned} \dot{\underline{\epsilon}}(\xi, \eta, \zeta) = & \dot{\underline{\epsilon}}(\mathcal{Q}) + \dot{\underline{\epsilon}}_{,\xi}(\mathcal{Q})\xi + \dot{\underline{\epsilon}}_{,\eta}(\mathcal{Q})\eta + \dot{\underline{\epsilon}}_{,\zeta}(\mathcal{Q})\zeta + \\ & 2\dot{\underline{\epsilon}}_{,\xi\eta}(\mathcal{Q})\xi\eta + 2\dot{\underline{\epsilon}}_{,\eta\zeta}(\mathcal{Q})\eta\zeta + 2\dot{\underline{\epsilon}}_{,\zeta\xi}(\mathcal{Q})\zeta\xi \end{aligned} \quad (21)$$

or

$$\dot{\underline{\epsilon}} = \sum_{a=1}^{NN} \bar{B}_a(\xi, \eta, \zeta) \underline{v}_a \quad (22)$$

where

$$\begin{aligned} \bar{B}_a(\xi, \eta, \zeta) = & B_a(Q) + B_{a,\xi}(Q)\xi + B_{a,\eta}(Q)\eta + B_{a,\zeta}(Q)\zeta + \\ & 2B_{a,\xi\eta}(Q)\xi\eta + 2B_{a,\eta\zeta}(Q)\eta\zeta + 2B_{a,\zeta\xi}(Q)\zeta\xi \end{aligned} \quad (23)$$

The first term of the right hand side of Equation (21) is the constant strain rates evaluated at the quadrature point,  $Q$ , and the remaining terms are linear and bilinear strain rate terms. After some tedious algebra, it can be shown that the first and second derivatives of  $B_a(Q)$  with respect to natural coordinates are given by:

$$b_{1,\xi} = \{N_{a,x\xi}\} = \frac{1}{8}[D_{12}\underline{\gamma}_1 + D_{13}\underline{\gamma}_2] \quad (24)$$

$$b_{2,\xi} = \{N_{a,y\xi}\} = \frac{1}{8}[D_{22}\underline{\gamma}_1 + D_{23}\underline{\gamma}_2] \quad (25)$$

$$b_{3,\xi} = \{N_{a,z\xi}\} = \frac{1}{8}[D_{32}\underline{\gamma}_1 + D_{33}\underline{\gamma}_2] \quad (26)$$

$$b_{1,\eta} = \{N_{a,x\eta}\} = \frac{1}{8}[D_{11}\underline{\gamma}_1 + D_{13}\underline{\gamma}_3] \quad (27)$$

$$b_{2,\eta} = \{N_{a,y\eta}\} = \frac{1}{8}[D_{21}\underline{\gamma}_1 + D_{23}\underline{\gamma}_3] \quad (28)$$

$$b_{3,\eta} = \{N_{a,z\eta}\} = \frac{1}{8}[D_{31}\underline{\gamma}_1 + D_{33}\underline{\gamma}_3] \quad (29)$$

$$b_{1,\zeta} = \{N_{a,x\zeta}\} = \frac{1}{8}[D_{11}\underline{\gamma}_2 + D_{12}\underline{\gamma}_3] \quad (30)$$

$$b_{2,\zeta} = \{N_{a,y\zeta}\} = \frac{1}{8}[D_{21}\underline{\gamma}_2 + D_{22}\underline{\gamma}_3] \quad (31)$$

$$b_{3,\zeta} = \{N_{a,z\zeta}\} = \frac{1}{8}[D_{31}\underline{\gamma}_2 + D_{32}\underline{\gamma}_3] \quad (32)$$

$$b_{1,\xi\eta} = \{N_{a,x\xi\eta}\} = \frac{1}{8}[D_{13}\underline{\gamma}_4 - (p_1^i \underline{x}_i) b_{i,\xi} - (r_1^i \underline{x}_i) b_{i,\eta}] \quad (33)$$

$$b_{2,\xi\eta} = \{N_{a,y\xi\eta}\} = \frac{1}{8}[D_{23}\underline{\gamma}_4 - (p_2^i \underline{x}_i) b_{i,\xi} - (r_2^i \underline{x}_i) b_{i,\eta}] \quad (34)$$

$$b_{3,\xi\eta} = (N_{a,z\xi\eta}) = \frac{1}{8}[D_{33}\gamma_4 - (p_3^i x_i)b_{i,\xi} - (r_3^i x_i)b_{i,\eta}] \quad (35)$$

$$b_{1,\eta\zeta} = (N_{a,x\eta\zeta}) = \frac{1}{8}[D_{11}\gamma_4 - (q_1^i x_i)b_{i,\eta} - (p_1^i x_i)b_{i,\zeta}] \quad (36)$$

$$b_{2,\eta\zeta} = (N_{a,y\eta\zeta}) = \frac{1}{8}[D_{21}\gamma_4 - (q_2^i x_i)b_{i,\eta} - (p_2^i x_i)b_{i,\zeta}] \quad (37)$$

$$b_{3,\eta\zeta} = (N_{a,x\eta\zeta}) = \frac{1}{8}[D_{31}\gamma_4 - (q_3^i x_i)b_{i,\eta} - (p_3^i x_i)b_{i,\zeta}] \quad (38)$$

$$b_{1,\xi\zeta} = (N_{a,x\xi\zeta}) = \frac{1}{8}[D_{12}\gamma_4 - (q_1^i x_i)b_{i,\xi} - (r_1^i x_i)b_{i,\zeta}] \quad (39)$$

$$b_{2,\xi\zeta} = (N_{a,y\xi\zeta}) = \frac{1}{8}[D_{22}\gamma_4 - (q_2^i x_i)b_{i,\xi} - (r_2^i x_i)b_{i,\zeta}] \quad (40)$$

$$b_{3,\xi\zeta} = (N_{a,z\xi\zeta}) = \frac{1}{8}[D_{32}\gamma_4 - (q_3^i x_i)b_{i,\xi} - (r_3^i x_i)b_{i,\zeta}] \quad (41)$$

where

$$p_i = D_{i1}h_1 + D_{i3}h_3 \quad (42)$$

$$q_i = D_{i1}h_2 + D_{i2}h_3 \quad (43)$$

$$r_i = D_{i2}h_1 + D_{i3}h_2 \quad (44)$$

The  $\gamma_\alpha$  in equations (24)-(41) are the stabilization vectors which span the improper null-space of  $\underline{B}(\underline{Q})$ . They are given by

$$\gamma_\alpha = h_\alpha - (h_\alpha^i x_i)h_i \quad (45)$$

where  $i$  is the summation index from 1 to 3.  $\gamma_\alpha$  are orthogonal to the linear displacement field and provide the proper stabilization for the element. It is also noted that the  $\gamma$ -stabilization element always meets the patch test while the  $h$ -stabilization element does not. Belytschko and coworkers [1-3] derive these vectors from consistency requirement.

To alleviate volumetric locking, we employ the ideas underlying selective/reduced integration (Hughes [10]).  $\bar{\underline{B}}_a(\xi, \eta, \zeta)$  is decomposed into two parts: the dilatational part and the deviatoric part. The dilatational part of gradient matrices are underintegrated and evaluated only at one quadrature point,  $\underline{Q}$ , to avoid volumetric locking:

$$\bar{\underline{B}}_a(\xi, \eta, \zeta) = \bar{\underline{B}}_a^{\text{dil}}(\underline{Q}) + \bar{\underline{B}}_a^{\text{dev}}(\xi, \eta, \zeta) \quad (46)$$

Expanding  $\bar{\underline{B}}_a^{\text{dev}}$  about the element center, equation (46) can be written as:



$$\begin{aligned}\bar{\mathbf{B}}_a(\xi, \eta, \zeta) = & \mathbf{B}_a(\mathbf{Q}) + \mathbf{B}_{a,\xi}^{\text{dev}}(\mathbf{Q})\xi + \mathbf{B}_{a,\eta}^{\text{dev}}(\mathbf{Q})\eta + \mathbf{B}_{a,\zeta}^{\text{dev}}(\mathbf{Q})\zeta + \\ & 2\mathbf{B}_{a,\xi\eta}^{\text{dev}}(\mathbf{Q})\xi\eta + 2\mathbf{B}_{a,\eta\zeta}^{\text{dev}}(\mathbf{Q})\eta\zeta + 2\mathbf{B}_{a,\zeta\xi}^{\text{dev}}(\mathbf{Q})\zeta\xi\end{aligned}\quad (47)$$

where  $\mathbf{B}_a(\mathbf{Q})$  are the one-point-quadrature gradient submatrices contributed from both the dilatational and deviatoric parts. The remaining terms on the right hand side of the above equation are the gradient submatrices corresponding to non-constant deviatoric strain rates. It is noted that the element using the gradient matrices as in equation (46) or (47) is properly underintegrated and exhibits no hourglass mode if the element internal energy is evaluated by using the multiple-point quadrature.

The element developed so far is not suitable to plate/shell analysis owing to the shear and membrane locking in thin structures. To remove shear locking, the gradient submatrices, corresponding to the assumed shear strain rates is written in an orthogonal corotational coordinate system rotating with the element (see Figure 2) as:

$$\begin{aligned}\bar{\mathbf{B}}_{xx}(\xi, \eta, \zeta) = & \mathbf{B}_{xx}(\mathbf{Q}) + \mathbf{B}_{xx,\xi}^{\text{dev}}(\mathbf{Q})\xi + \mathbf{B}_{xx,\eta}^{\text{dev}}(\mathbf{Q})\eta + \mathbf{B}_{xx,\zeta}^{\text{dev}}(\mathbf{Q})\zeta + \\ & 2\mathbf{B}_{xx,\xi\eta}^{\text{dev}}(\mathbf{Q})\xi\eta + 2\mathbf{B}_{xx,\eta\zeta}^{\text{dev}}(\mathbf{Q})\eta\zeta + 2\mathbf{B}_{xx,\zeta\xi}^{\text{dev}}(\mathbf{Q})\zeta\xi\end{aligned}\quad (48)$$

$$\begin{aligned}\bar{\mathbf{B}}_{yy}(\xi, \eta, \zeta) = & \mathbf{B}_{yy}(\mathbf{Q}) + \mathbf{B}_{yy,\xi}^{\text{dev}}(\mathbf{Q})\xi + \mathbf{B}_{yy,\eta}^{\text{dev}}(\mathbf{Q})\eta + \mathbf{B}_{yy,\zeta}^{\text{dev}}(\mathbf{Q})\zeta + \\ & 2\mathbf{B}_{yy,\xi\eta}^{\text{dev}}(\mathbf{Q})\xi\eta + 2\mathbf{B}_{yy,\eta\zeta}^{\text{dev}}(\mathbf{Q})\eta\zeta + 2\mathbf{B}_{yy,\zeta\xi}^{\text{dev}}(\mathbf{Q})\zeta\xi\end{aligned}\quad (49)$$

$$\begin{aligned}\bar{\mathbf{B}}_{zz}(\xi, \eta, \zeta) = & \mathbf{B}_{zz}(\mathbf{Q}) + \mathbf{B}_{zz,\xi}^{\text{dev}}(\mathbf{Q})\xi + \mathbf{B}_{zz,\eta}^{\text{dev}}(\mathbf{Q})\eta + \mathbf{B}_{zz,\zeta}^{\text{dev}}(\mathbf{Q})\zeta + \\ & 2\mathbf{B}_{zz,\xi\eta}^{\text{dev}}(\mathbf{Q})\xi\eta + 2\mathbf{B}_{zz,\eta\zeta}^{\text{dev}}(\mathbf{Q})\eta\zeta + 2\mathbf{B}_{zz,\zeta\xi}^{\text{dev}}(\mathbf{Q})\zeta\xi\end{aligned}\quad (50)$$

$$\bar{\mathbf{B}}_{xy}(\xi, \eta, \zeta) = \mathbf{B}_{xy}(\mathbf{Q}) + \mathbf{B}_{xy,\zeta}^{\text{dev}}(\mathbf{Q})\zeta\quad (51)$$

$$\bar{\mathbf{B}}_{yz}(\xi, \eta, \zeta) = \mathbf{B}_{yz}(\mathbf{Q}) + \mathbf{B}_{yz,\xi}^{\text{dev}}(\mathbf{Q})\xi\quad (52)$$

$$\bar{\mathbf{B}}_{zx}(\xi, \eta, \zeta) = \mathbf{B}_{zx}(\mathbf{Q}) + \mathbf{B}_{zx,\eta}^{\text{dev}}(\mathbf{Q})\eta\quad (53)$$

where  $\bar{\mathbf{B}}_{xx}$ ,  $\bar{\mathbf{B}}_{yy}$ ,  $\bar{\mathbf{B}}_{zz}$ ,  $\bar{\mathbf{B}}_{xy}$ ,  $\bar{\mathbf{B}}_{yz}$  and  $\bar{\mathbf{B}}_{zx}$  are the gradient submatrices corresponding to strain rates  $\dot{\epsilon}_{xx}$ ,  $\dot{\epsilon}_{yy}$ ,  $\dot{\epsilon}_{zz}$ ,  $\dot{\epsilon}_{xy}$ ,  $\dot{\epsilon}_{yz}$  and  $\dot{\epsilon}_{zx}$ , respectively. Here, only one non-constant term is used for each shear strain rate component such that the modes causing shear locking are removed. The normal strain rates keep all non-constant terms given in equation (47). In  $\mathbf{B}^{\text{dev}}(\mathbf{Q})$ , only those terms corresponding to a parallelepiped element are used for stabilization.

To detect plastic fronts in the mesh during loading and unloading more accurately for elastic-plastic large deformation problems, we propose to use a four-point-quadrature scheme instead of the one-point-quadrature. The element internal force vector is evaluated at the four integration points located as follows:

$$\begin{aligned}
\text{Point 1: } & \left( +\frac{1}{\sqrt{3}}, +\frac{1}{\sqrt{3}}, +\frac{1}{\sqrt{3}} \right); & \text{Point 2: } & \left( -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, +\frac{1}{\sqrt{3}} \right) \\
\text{Point 3: } & \left( -\frac{1}{\sqrt{3}}, +\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}} \right); & \text{Point 4: } & \left( +\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}} \right)
\end{aligned} \tag{54}$$

This element exhibits no hourglass mode and is rank sufficient.

By assuming that the Jacobian is a constant, one quarter of the element volume, the element internal force vector can be integrated as follows:

$$\mathbf{f}^{\text{int}} = \sum_{k=1}^4 \frac{V}{4} \bar{\mathbf{B}}^T(\underline{\xi}_k) \mathbf{T}(\underline{\xi}_k) \tag{55}$$

where  $\underline{\xi}_k$  denotes the natural coordinates of the integration point  $k$  and  $V$  is the element volume.

The element internal force vector can be rearranged in the form:

$$\mathbf{f}^{\text{int}} = \mathbf{f}_4^{\text{int}} + \mathbf{f}_{\text{stab}}^{\text{int}} \tag{56}$$

where  $\mathbf{f}_4^{\text{int}}$  and  $\mathbf{f}_{\text{stab}}^{\text{int}}$  are the internal force vectors resulting from the one-point quadrature and the stabilization procedure, respectively. They are given by:

$$\mathbf{f}_4^{\text{int}} = \sum_{k=1}^4 \frac{V}{4} \begin{bmatrix} \tau_{11}(\underline{\xi}_k) b_1 + \tau_{12}(\underline{\xi}_k) b_2 + \tau_{13}(\underline{\xi}_k) b_3 \\ \tau_{21}(\underline{\xi}_k) b_1 + \tau_{22}(\underline{\xi}_k) b_2 + \tau_{23}(\underline{\xi}_k) b_3 \\ \tau_{31}(\underline{\xi}_k) b_1 + \tau_{32}(\underline{\xi}_k) b_2 + \tau_{33}(\underline{\xi}_k) b_3 \end{bmatrix} \tag{57}$$

and

$$\mathbf{f}_{\text{stab}}^{\text{int}} = \sum_{k=1}^4 \frac{V}{32} \begin{bmatrix} \underline{g}_1(\underline{\xi}_k) \tau_{11}^{\text{dev}}(\underline{\xi}_k) + \underline{g}_4(\underline{\xi}_k) \tau_{12}(\underline{\xi}_k) + \underline{g}_5(\underline{\xi}_k) \tau_{31}(\underline{\xi}_k) \\ \underline{g}_2(\underline{\xi}_k) \tau_{22}^{\text{dev}}(\underline{\xi}_k) + \underline{g}_6(\underline{\xi}_k) \tau_{12}(\underline{\xi}_k) + \underline{g}_7(\underline{\xi}_k) \tau_{23}(\underline{\xi}_k) \\ \underline{g}_3(\underline{\xi}_k) \tau_{33}^{\text{dev}}(\underline{\xi}_k) + \underline{g}_8(\underline{\xi}_k) \tau_{23}(\underline{\xi}_k) + \underline{g}_9(\underline{\xi}_k) \tau_{31}(\underline{\xi}_k) \end{bmatrix} \tag{58}$$

where the superscript, dev, denotes the deviatoric part of the stress ( $\tau_{ij}^{\text{dev}} = \tau_{ij} - \frac{1}{3} \tau_{kk} \delta_{ij}$ ) and the other quantities are given by

$$\underline{g}_1(\underline{\xi}) = D_{11}(\eta \gamma_1 + \zeta \gamma_2 + 2\eta \zeta \gamma_4) \tag{59}$$

$$\underline{g}_2(\underline{\xi}) = D_{22}(\xi \gamma_1 + \zeta \gamma_3 + 2\xi \zeta \gamma_4) \tag{60}$$

$$\underline{g}_3(\underline{\xi}) = D_{33}(\xi \gamma_2 + \eta \gamma_3 + 2\xi \eta \gamma_4) \tag{61}$$

$$\underline{g}_4(\underline{\xi}) = D_{22}\underline{\zeta}\underline{\gamma}_3 \quad (62)$$

$$\underline{g}_5(\underline{\xi}) = D_{33}\underline{\eta}\underline{\gamma}_3 \quad (63)$$

$$\underline{g}_6(\underline{\xi}) = D_{11}\underline{\zeta}\underline{\gamma}_2 \quad (64)$$

$$\underline{g}_7(\underline{\xi}) = D_{33}\underline{\xi}\underline{\gamma}_2 \quad (65)$$

$$\underline{g}_8(\underline{\xi}) = D_{22}\underline{\xi}\underline{\gamma}_1 \quad (66)$$

$$\underline{g}_9(\underline{\xi}) = D_{11}\underline{\eta}\underline{\gamma}_1 \quad (67)$$

It is noted that the elements developed above can not pass the patch test because with one-point quadrature the element internal forces are not properly evaluated if the elements are skewed (Belytschko and Bindeman [7]). To remedy this drawback,  $\underline{B}_a(Q)$  are replaced by the uniform gradient matrices,  $\tilde{\underline{B}}_a$ , defined by Belytschko *et al.* [1,2]:

$$\tilde{\underline{B}}_a = \frac{1}{V_e} \int_{\Omega} \underline{B}_a(\underline{\xi}, \underline{\eta}, \underline{\zeta}) dV \quad (68)$$

where  $V_e$  is the element volume. Similarly, equation (4) is modified as

$$\tilde{\underline{B}}_a(Q) = \begin{bmatrix} \tilde{b}_{1a} \\ \tilde{b}_{2a} \\ \tilde{b}_{3a} \end{bmatrix} \quad (69)$$

and the stabilization vectors are redefined as:

$$\tilde{\underline{\gamma}}_\alpha = \underline{h}_\alpha - (\underline{h}_\alpha^T \underline{X}_i) \tilde{\underline{b}}_i \quad (70)$$

which span the proper null-space. Since, the element internal force vector can be evaluated exactly when the element is subjected to a constant strain rate field, the use of the uniform gradient matrices  $\tilde{\underline{B}}_a$  leads a new four-quadrature-point element, the NUHEXIN-4 element, which passes the patch test.

The three dimensional ASQBI (assumed strain quintessential bending incompressible) element is developed by Belytschko and Bindeman [7]. The gradient matrix,  $\hat{\underline{B}}$ , is given by

$$\begin{bmatrix} \hat{E}_{xx}(\xi, \eta, \zeta) \\ \hat{E}_{yy}(\xi, \eta, \zeta) \\ \hat{E}_{zz}(\xi, \eta, \zeta) \\ \hat{E}_{xy}(\xi, \eta, \zeta) \\ \hat{E}_{yz}(\xi, \eta, \zeta) \\ \hat{E}_{zx}(\xi, \eta, \zeta) \end{bmatrix} = \begin{bmatrix} \tilde{b}_1 & 0 & 0 \\ 0 & \tilde{b}_2 & 0 \\ 0 & 0 & \tilde{b}_3 \\ \tilde{b}_2 & \tilde{b}_1 & 0 \\ 0 & \tilde{b}_3 & \tilde{b}_2 \\ \tilde{b}_3 & 0 & \tilde{b}_1 \end{bmatrix} +$$

$$\begin{bmatrix} h_{1,x}\tilde{\gamma}_1 + h_{2,x}\tilde{\gamma}_2 + h_{4,x}\tilde{\gamma}_4 & -\bar{v}h_{1,x}\tilde{\gamma}_1 - \bar{v}h_{4,x}\tilde{\gamma}_4 & -\bar{v}h_{2,x}\tilde{\gamma}_2 - \bar{v}h_{4,x}\tilde{\gamma}_4 \\ -\bar{v}h_{1,x}\tilde{\gamma}_1 - \bar{v}h_{4,x}\tilde{\gamma}_4 & h_{1,y}\tilde{\gamma}_1 + h_{3,y}\tilde{\gamma}_3 + h_{4,y}\tilde{\gamma}_4 & -\bar{v}h_{3,x}\tilde{\gamma}_3 - \bar{v}h_{4,x}\tilde{\gamma}_4 \\ -\bar{v}h_{2,x}\tilde{\gamma}_2 - \bar{v}h_{4,x}\tilde{\gamma}_4 & -\bar{v}h_{3,x}\tilde{\gamma}_3 - \bar{v}h_{4,x}\tilde{\gamma}_4 & h_{2,z}\tilde{\gamma}_2 + h_{3,z}\tilde{\gamma}_3 + h_{4,z}\tilde{\gamma}_4 \\ h_{3,y}\tilde{\gamma}_3 & h_{2,x}\tilde{\gamma}_2 & 0 \\ 0 & h_{2,z}\tilde{\gamma}_2 & h_{1,y}\tilde{\gamma}_1 \\ h_{3,z}\tilde{\gamma}_3 & 0 & h_{1,x}\tilde{\gamma}_1 \end{bmatrix} \quad (71)$$

where

$$h_1 = \xi\eta \quad h_2 = \xi\zeta \quad h_3 = \eta\zeta \quad h_4 = \xi\eta\zeta \quad (72)$$

;  $v$  is the Poisson's ratio;  $\bar{v} = \frac{v}{1-v}$ .

To compare the NUHEXIN-4 and ASQBI elements, let us write out the gradient matrix corresponding to the NUHEXIN-4 element can be shown to be:

$$\begin{bmatrix} \bar{E}_{xx}(\xi, \eta, \zeta) \\ \bar{E}_{yy}(\xi, \eta, \zeta) \\ \bar{E}_{zz}(\xi, \eta, \zeta) \\ \bar{E}_{xy}(\xi, \eta, \zeta) \\ \bar{E}_{yz}(\xi, \eta, \zeta) \\ \bar{E}_{zx}(\xi, \eta, \zeta) \end{bmatrix} = \begin{bmatrix} \tilde{b}_1 & 0 & 0 \\ 0 & \tilde{b}_2 & 0 \\ 0 & 0 & \tilde{b}_3 \\ \tilde{b}_2 & \tilde{b}_1 & 0 \\ 0 & \tilde{b}_3 & \tilde{b}_2 \\ \tilde{b}_3 & 0 & \tilde{b}_1 \end{bmatrix} +$$

$$\frac{1}{8} \begin{bmatrix} \frac{2}{3}D_{11}(\eta\bar{\gamma}_1+\zeta\bar{\gamma}_2+2\eta\zeta\bar{\gamma}_4) & \frac{1}{3}D_{22}(\xi\bar{\gamma}_1+\zeta\bar{\gamma}_3+2\xi\zeta\bar{\gamma}_4) & \frac{1}{3}D_{33}(\xi\bar{\gamma}_2+\eta\bar{\gamma}_3+2\xi\eta\bar{\gamma}_4) \\ \frac{1}{3}D_{11}(\eta\bar{\gamma}_1+\zeta\bar{\gamma}_2+2\eta\zeta\bar{\gamma}_4) & \frac{2}{3}D_{22}(\xi\bar{\gamma}_1+\zeta\bar{\gamma}_3+2\xi\zeta\bar{\gamma}_4) & \frac{1}{3}D_{33}(\xi\bar{\gamma}_2+\eta\bar{\gamma}_3+2\xi\eta\bar{\gamma}_4) \\ \frac{1}{3}D_{11}(\eta\bar{\gamma}_1+\zeta\bar{\gamma}_2+2\eta\zeta\bar{\gamma}_4) & \frac{1}{3}D_{22}(\xi\bar{\gamma}_1+\zeta\bar{\gamma}_3+2\xi\zeta\bar{\gamma}_4) & \frac{2}{3}D_{33}(\xi\bar{\gamma}_2+\eta\bar{\gamma}_3+2\xi\eta\bar{\gamma}_4) \\ D_{22}\zeta\bar{\gamma}_3 & D_{11}\zeta\bar{\gamma}_2 & 0 \\ 0 & D_{33}\xi\bar{\gamma}_2 & D_{22}\xi\bar{\gamma}_1 \\ D_{33}\eta\bar{\gamma}_3 & 0 & D_{11}\eta\bar{\gamma}_1 \end{bmatrix} \quad (73)$$

It can be seen from Equations (71) and (73) that both gradient matrices have the same number and locations of zero and non-zero components. Both matrices including all the stabilization vectors to suppress the hourglass modes and the resulting stiffness matrices are rank sufficient. It is interesting to point out that those components corresponding to the assumed shear strain rates have the same forms but with different coefficients. It is believed that the Belytschko-Bindeman element is computationally more involved since the gradient matrix is written in terms of the spatial derivatives of  $\xi\eta$ ,  $\eta\zeta$ ,  $\zeta\xi$  and  $\xi\eta\zeta$ ; however, the gradient matrix of NUHEXIN-4 element is written in terms of  $\xi$ ,  $\eta$  and  $\zeta$  explicitly so that the stiffness matrix can be easily obtained through numerical integration.

**EXAMPLES.** In this section, a variety of problems including beams, plates and shells are studied to investigate the performance of the proposed NUHEXIN-4 element. Besides this element, runs are also made by using another element called NUHEX-4, in which the virtual work corresponding to the dilatation part is evaluated by using a four-point-quadrature scheme instead of the one-point-quadrature used in the NUHEXIN-4 element. Since the NUHEXIN-4 element is mainly proposed to be used in sheet metal forming analysis, the applicability of NUHEXIN-4 element to problems of thin structures is also studied by solving the standard test problems including the twisted beam, pinched cylinder, Scordelis-Lo roof and hemispherical shell, which are proposed by MacNeal-Harder [11] and Belytschko *et al.* [12].

**Clamped Twisted Beam.** In this example, a cantilever twisted beam subjected to a uniform shear force ( $F=1.0$ ) at the free end is analyzed. The problem statement is shown in Figure 3. The computed displacement the free end  $w$  is compared to the analytical solution in Table 1. Unlike the curved beam problem, the NUHEX-4 element performs better.

**Simply Supported Plate.** In this example, a simply supported plate subjected to a concentrated load at the center is analyzed. The problem statement is shown in Figure 4. Due to symmetry, only one quarter of the plate is modelled. Two cases are studied in this example: regular mesh and irregular mesh. The computed central displacement  $w$  is compared to the analytical solution in Table 2 and 3. Both elements compare well with the analytical solution for the uniform mesh; whereas the NUHEXIN-4 performs much better for the irregular mesh.

**Pinched Cylinder.** Figure 5. shows the pinched cylinder subjected to concentrated loads. Two cases are studied in this example. In the first case, it is assumed that the both ends of the cylinder are free. In the second case, it is assumed that the both end of the cylinder are covered with rigid diaphragms so that only the displacement in the axial direction is allowed at the ends. Due to symmetry, only one quarter of the cylinder is modelled. The computed central displacements are compared to the analytical solution in Table 4 and 5. The NUHEXIN-4 element performs much better than the NUHEX-4 element for the pinched cylinder with diaphragms and both elements perform approximately the same for that without diaphragm.

**Scordelis-Lo Roof.** Figure 6. shows the Scordelis-Lo roof subjected to its own weight. Due to symmetry, only one quarter of the roof is modelled. It is assumed that the both ends of the roof are covered with rigid diaphragms so that only the displacement in the axial direction is allowed at the ends. The computed central edge displacement  $w$  is compared to the analytical solution in Table 6. As can be seen, the NUHEX-4 performs better.

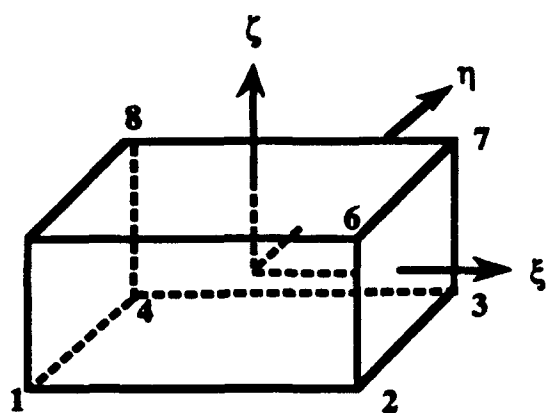
**Hemispherical Shell.** Figure 7 shows the hemisphere shell subjected to antisymmetrical concentrated loads at its bottom ends. Due to symmetry, only one quarter of the hemispherical shell is modelled. The computed radial deflection  $w$  is compared to the analytical solution in Table 7. Due to the double curvature of the shell,  $48 \times 48$  elements are required to model the geometry correctly with NUHEXIN-4 element.

**CONCLUSIONS.** In this paper, the multiple-quadrature-point eight-node hexahedral elements are developed. The emphasis has been placed on the removal of shear and volumetric locking and the suppression of spurious singular modes. The stabilization operations are obtained simply by taking the partial derivatives of the generalized strain rate vector with respect to the natural coordinates. The resulting element stiffness matrices can be explicitly expressed in terms of natural coordinates so that they are easier to compute than those involving spatial derivatives. The performances of the proposed elements are studied by solving beam, plate and shell problems. It is demonstrated that the solutions are satisfactory.

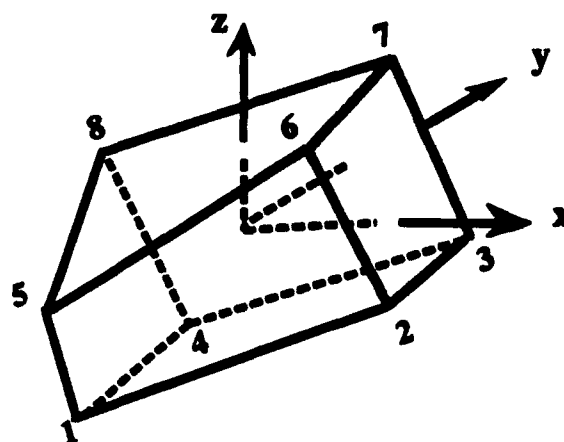
Finally, it should be pointed out that the performances of the NUHEX-4 and NUHEXIN-4 elements are approximately the same in the compressible problems; however NUHEXIN-4 should be used in the elastic-plastic and incompressible problems since the NUHEX-4 element suffers from volumetric locking.

## REFERENCES

1. Flanagan, D.P. and Belytschko, T., "A uniform strain hexahedron and quadrilateral with orthogonal hourglass control," *Int. J. Num. Methods Engng.* 17, pp. 679-706, (1981).
2. Belytschko, T., "Correction of Article by Flanagan, D.P. and Belytschko, T.," *Int. J. Num. Methods Engng.* 19, pp. 467-468, (1983).
3. Belytschko, T., Ong, J. S.-J., Liu, W.K. and Kennedy, J.M., "Hourglass control in linear and nonlinear problems", *Comput. Methods Appl. Mech. Eng.* 43, pp. 251-276, (1984).
4. Liu, W.K. and Belytschko, T., "Efficient linear and nonlinear heat conduction with a quadrilateral element," *Int. J. Num. Methods Engng.* 20, pp. 931-948, (1984).
5. Belytschko, T. and Bachrach, W.E., "Efficient implementation of quadrilaterals with high coarse-mesh accuracy," *Comp. Methods Appl. Mech. Eng.* 54, pp. 279-301, (1986).
6. Belytschko, T. and Bindeman, L.P., "Assumed strain stabilization of the 4-node quadrilateral with 1-point quadrature for nonlinear problems," *Comput. Methods Appl. Mech. Eng.* 88, pp. 311-340, (1991).
7. Belytschko, T. and Bindeman, L.P., "Assumed strain stabilization of the eight node hexahedral element," to be published.
8. Liu, W.K., Ong, J.S.-J., and Uras, R.A., 'Finite element stabilization matrices-A unification approach,' *Comput. Methods Appl. Mech. Eng.* 53, pp. 13-46, (1985).
9. Koh, B.C. and Kikuchi, N., "New Improved Hourglass Control for Bilinear and Trilinear Elements in Anisotropic Linear Elasticity," *Comput. Meths. Appl. Mech. Engng.* 65, pp. 1-46, (1987).
10. Hughes, T.J.R., "Generalization of selective integration procedures to anisotropic and nonlinear media", *Internat. Numer. Meths. Engrg.*, Vol. 15, pp. 1413-1418, (1980).
11. MacNeal R.H. and Harder R.L., "A proposed standard set of problems to test finite element accuracy," *Finite Elements Anal. Des.*, Vol. 11, pp. 3-20, (1985).
12. Belytschko, T., Stolarski, H. and Carpenter, N., "A  $C_0$  triangular plate element with one-point quadrature," *Int. J. Num. Methods Engng* 20, pp. 787-802, (1984).



Referential coordinat system



Physical coordinat system

Figure 1. An eight-node hexahedral element in referential and physical coordinate systems.

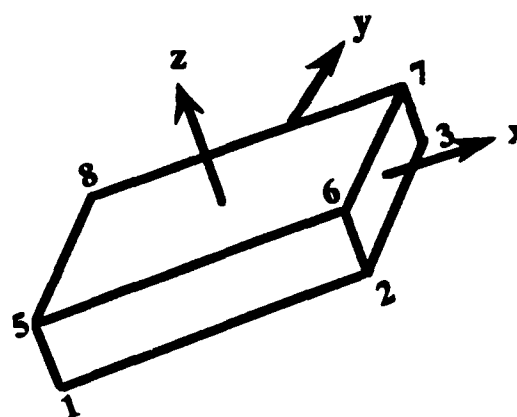
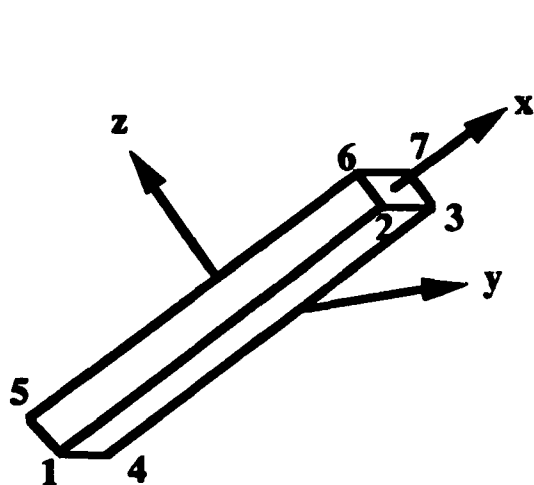
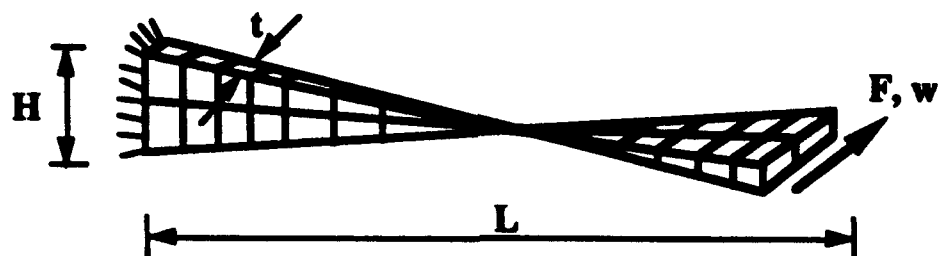


Figure 2. Corotational coordinate systems in three dimensions



$$E = 2.9 \times 10^7$$

$$\nu = 0.22$$

$$L = 12$$

$$H = 1.1$$

$$t = 0.32$$

$$F = 1.0$$

$$\text{Twist} = 90^\circ$$

Figure 3. Problem statement of the clamped twisted beam

Analytical solution  $w = 0.005424$

Mesh Element	6 x 1 x 1	12 x 2 x 2	24 x 4 x 4
NUHEX-4	3.278	1.081	1.003
NUHEXIN-4	11.299	1.157	1.026

Table. 1 Normalized displacement of the twisted beam



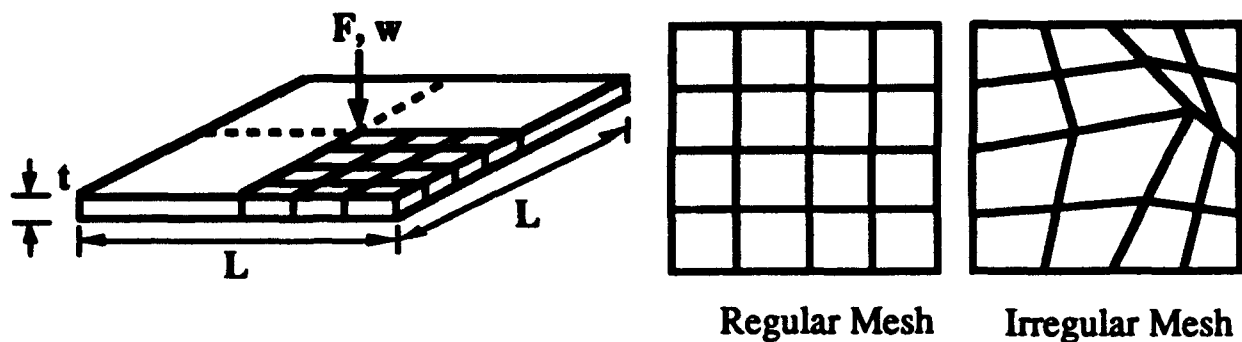


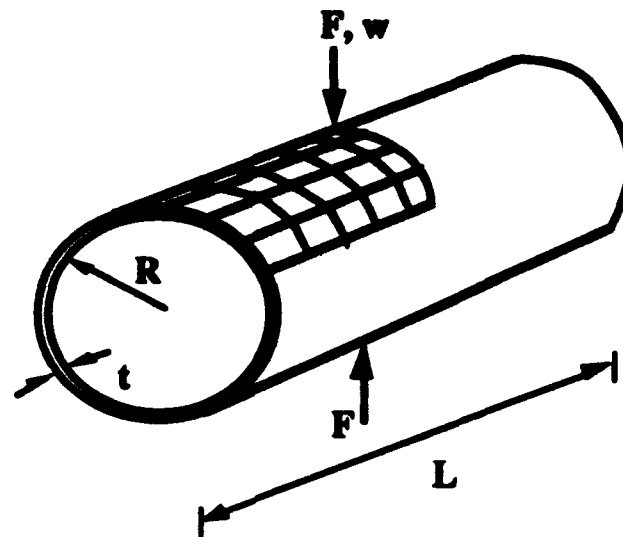
Figure 4. Problem statement of the simply supported plate

Analytical solution $w = 0.021138$			
Mesh Element	4 x 4 x 2	8 x 8 x 4	16 x 16 x 4
NUHEX-4	0.943	0.985	0.987
NUHEXIN-4	1.151	1.034	1.036

Table. 2 Normalized displacement of the simply supported plate with regular meshes

Analytical solution $w = 0.021138$			
Mesh Element	4 x 4 x 2	8 x 8 x 4	16 x 16 x 4
NUHEX-4	0.611	0.848	0.939
NUHEXIN-4	0.704	0.867	0.984

Table. 3 Normalized displacement of the simple supported plate with irregular meshes



First Case (with diaphragms)

$E = 1.05 \times 10^6$   $\nu = 0.3125$   
 $L = 10.35$   $t = 0.094$   
 $F = 100.0$   $R = 4.953$

Second Case (without diaphragm)

$E = 3 \times 10^6$   $\nu = 0.3$   
 $L = 600.0$   $t = 3.0$   
 $F = 1.0$   $R = 300.0$

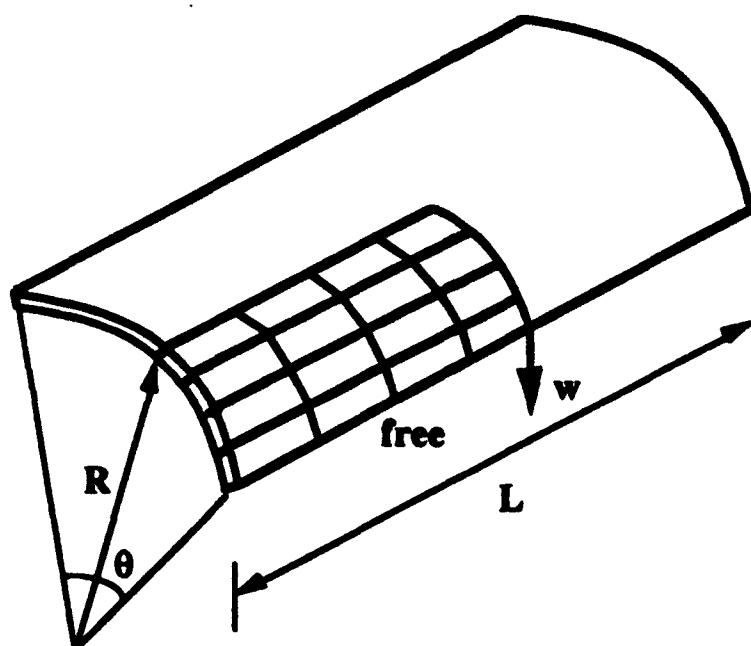
Figure 5. Problem statement of the pinched cylinder

Analytical solution $w = 0.1137$			
Mesh Element	10 x 10 x 2	15 x 15 x 4	20 x 20 x 4
NUHEX-4	0.927	0.998	1.005
NUHEXIN-4	1.145	1.050	1.055

Table. 4 Normalized displacement of the pinched cylinder without diaphragm

Analytical solution $w = 0.000018248$			
Mesh Element	10 x 10 x 2	15 x 15 x 4	20 x 20 x 4
NUHEX-4	0.633	0.870	0.936
NUHEXIN-4	0.811	0.934	0.980

Table. 5 Normalized displacement of the pinched cylinder with diaphragms



$$E = 4.32 \times 10^8$$

$$\nu = 0.0$$

$$L = 50.0$$

$$R = 25.0$$

$$t = 0.25$$

$$\text{Gravity} = 360.0/\text{per volume}$$

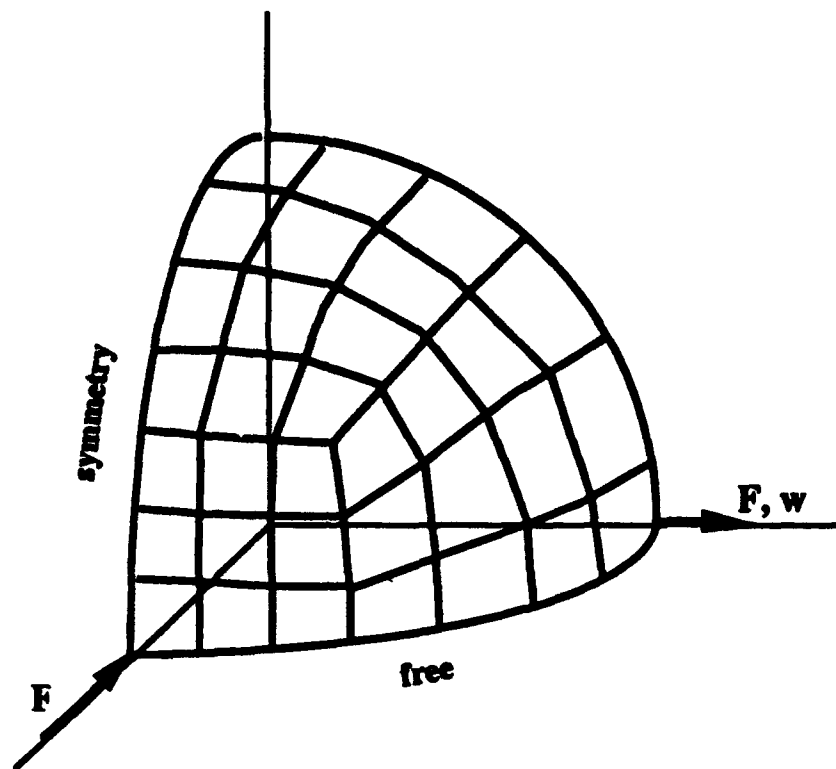
$$\theta = 80^\circ$$

Figure 6. Problem statement of the Scordelis-Lo roof

Analytical solution  $w = 0.3024$

Mesh Element	8 x 8 x 1	16 x 16 x 1	32 x 32 x 1
NUHEX-4	1.016	1.011	1.010
NUHEXIN-4	1.162	1.144	1.140

Table. 6 Normalized displacement of the Scordelis-Lo roof



$$E = 6.825 \times 10^7 \quad \nu = 0.3$$

$$\text{Radius} = 10.0 \quad \text{thickness} = 0.4$$

$$F = 1.0$$

Figure 7. Problem statement of the hemispherical shell

Analytical solution $w = 0.0924$			
Element \ Mesh*	16 x 1	32 x 1	48 x 1
NUHEX-4	0.615	0.862	0.954
NUHEXIN-4	0.638	0.891	0.984

( \*elements /side x elements/thickness )

Table. 7 Normalized displacement of the hemispherical shell

# THREE-DIMENSIONAL FINITE ELEMENT MODEL GENERATION AND RESPONSE SIMULATION OF AN ARMORED VEHICLE

Aaron D. Gupta  
Joseph M. Santiago  
Henry L. Wisniewski

U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland

## ABSTRACT

Light combat vehicles are playing an important support role for both troops and other heavily armored combat vehicles. As such, they have a much greater risk than in previous roles of being subjected to transient loads such as impact and overpressure loads. Propagation of ballistic shock from an impacted region to the critical locations and attachment points for secondary systems can cause damage and misalignment to sensitive equipments contributing to malfunction and reduction of vehicle performance. Accuracy of determination of dynamic response of these vehicles is directly dependent on the degree of refinement of the generated model and how well the model incorporates the essential features of the vehicle and correlates to its important characteristics without being overburdened by non-essential details. Additionally, response of nonlinear components of the vehicle in high frequency regime may influence the overall global response of the vehicle. As a result, hatch openings and access door cutouts with unsymmetric locations may have to be incorporated in the finite element model to allow fair comparison with first order experiments involving a stripped vehicle hull. The current study is an attempt to assess the influence of multiple rectangular cutouts on the overall transient response of a vehicle hull subjected to a side-on impact load.

## 1. INTRODUCTION

The analysis of the dynamic response of complex systems involving structural assemblies and components has become a subject of considerable research because of its practical significance in the evaluation of structural integrity when subjected to transient loads [1-4]. The behavior of combat vehicles subjected to dynamic loads is of particular interest to the Army because of the need to ensure survivability and minimize degradation of performance in both primary and secondary systems. An area of critical concern in a complex system such as a combat vehicle or a personnel carrier is the propagation of shock from an impact point on the vehicle hull to the location of the driver and the other personnel in the crew compartment and the attachment points for the optical equipments, such as the gunsights and periscopes, as well as electronically sensitive command and control devices. The fail-

ure of equipments and structures in such a system resulting from shock propagation and vibrations can render the system ineffective and vulnerable to enemy attack, leading to life threatening situations. Detailed dynamic response analysis of such systems are essential in order to develop nondestructive evaluation methods to ensure survivability of critical Army equipments and hardware.

Armored vehicles must survive the shock resulting from non-penetrating projectile impact and be able to retain fighting capability after this impact. Non-standard shock measurement techniques were employed by W. Scott Walton [5,6] to determine ballistic shock protection requirements for armored combat vehicles. Target loading histories of thin plates were developed using the EPIC-2 Lagrangian code by E.F. Quigley [7,8] who used the data as input to the ADINA nonlinear finite element analysis code to predict the ballistic shock environments for normal impacted, axisymmetric targets.

The response of structures at critical locations is directly dependent on ballistic shock propagation from the point of load application to the point of attachment of secondary sensitive equipments when subjected to high frequency loads, such as projectile impact and close-in blast overpressure. To evaluate the post-impact response and performance of the system, both frequency domain and time domain response of the system should be conducted. However, in the current investigation a finite element approach was employed to obtain responses at specific locations of the model. The time domain responses were converted to frequency domain responses using a conversion code based on [9].

## 2. PROBLEM CONFIGURATION

The particular vehicle selected for this simulation is the Armoured Personnel Carrier (APC) designated as APC M113. Overall specifications are available in Jane's World Armoured Fighting Vehicles [10]. Detailed dimensions and geometry description of the vehicle was obtained from field measurements and from [11]. Overall length, width and height of the APC are 4.863m, 2.686m and 2.5m respectively while the height of the hull top is only 1.828m. Unloaded weight of the vehicle is 9702 kg while the fully loaded vehicle weighs 11156 kg. Ground clearance for the loaded vehicle is 0.406m. Details of the track assembly were completely left out since the hull was modeled without tracks to avoid complexities for the 3D model generation. However, subsequent study of dynamic response due to transient loads included artificial boundary conditions imposed on the corner nodes of the bottom floor to avoid large scale sliding, rotation and overturning. A pictorial view of the entire vehicle assembly is shown in Figure 1.

The critical locations for the driver's and the commander's seats as well as the attachment points for shock sensitive control panels and periscopes were measured from an actual vehicle at the Combat Support and Test Activity (CSTA) at the Aberdeen Proving Ground and the corresponding nodal locations on the vehicle hull were determined for the purpose of response prediction since the simplified model did not include any internal components. Additionally, the driver's hatch, cargo hatch as well as the hatch opening for the commander's cupola were included as simulated rectangular cutouts in the generic model on the top hull surface in Figure 2 to assess the influence of multiple cutouts on the overall response of the model as a first step towards simulation of a basic vehicle hull without any tight fitting

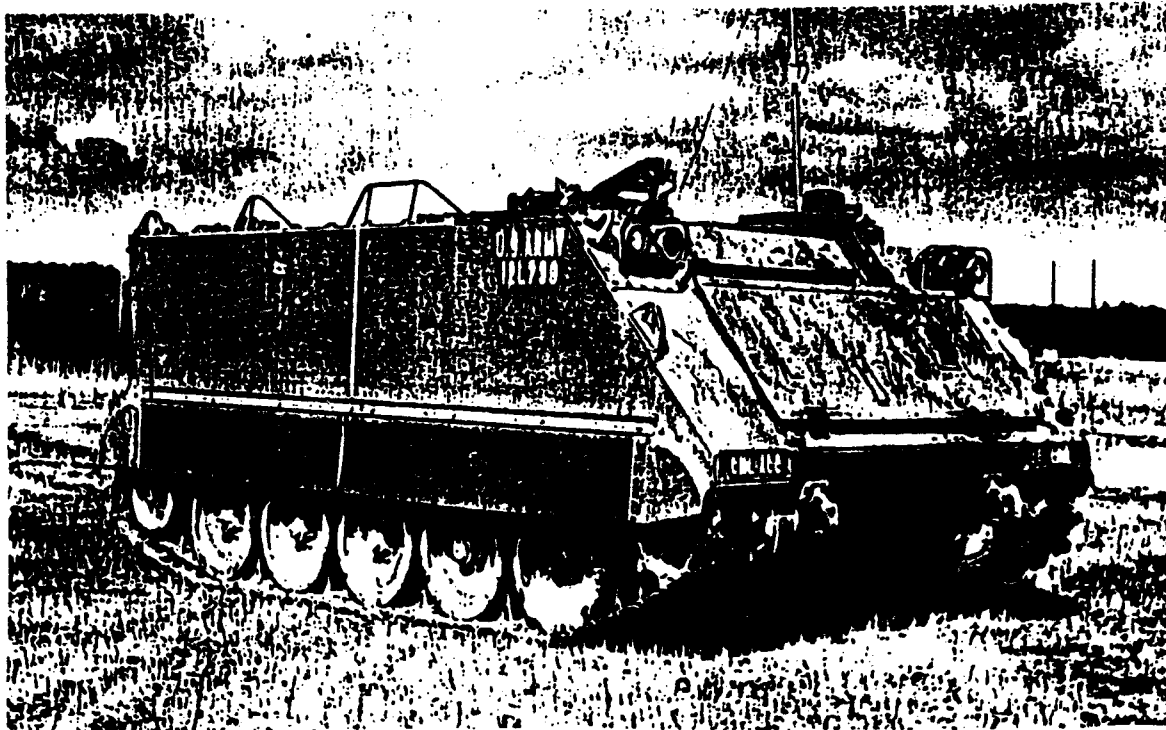


Figure 1. US Army Armored Personnel Carrier, M113A1

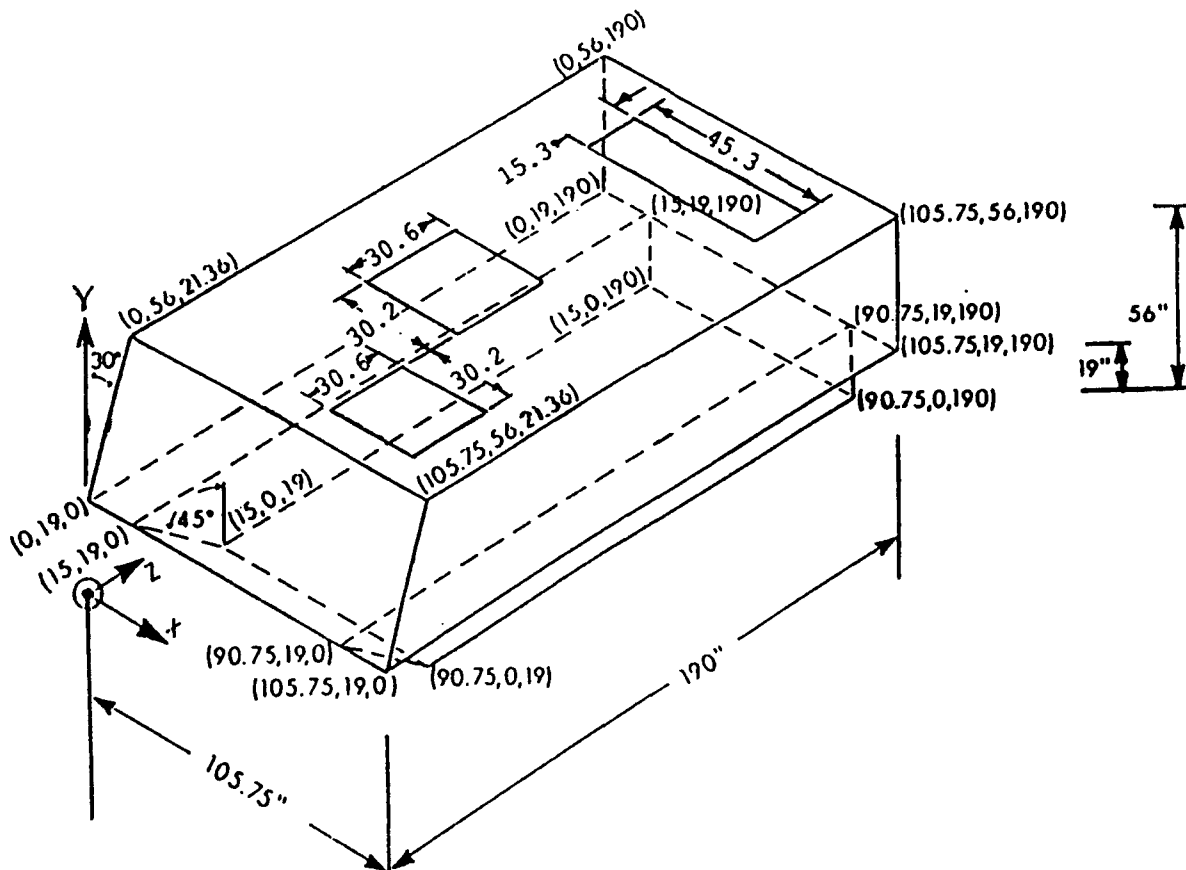


Figure 2. Generic APC M113 hull model with cutouts.  
(All dimensions given in inches)

components prior to analysis of a full-up vehicle. Although the actual openings in the vehicle hull were circular shaped except the cargo hatch and the rear door, it was decided to use approximate rectangular cutouts for all openings because these could be quickly generated through elimination of a cluster of rectangular elements in appropriate locations.

### 3. LOAD ESTIMATION

The modeshapes and eigenfrequencies analyses of the hull structure in ADINA [12] do not require transient loading of the vehicle. However, simulation of forced response of the vehicle require imposition of impact and pressure loads on one side of the vehicle at a specific location which will likely result in shock propagation and damage to primary as well as critical secondary systems inside the vehicle.

Impact load due to side-on impact of a projectile approximately 1.83 m long weighing approximately 6.8 kg and travelling at .914 km/s is calculated assuming that the rod continues to erode at a constant rate determined by the initial velocity of the rod until it is fully consumed and the total momentum of the rod is imparted as impulse to the side of the vehicle. The impact load is imposed as a concentrated load on a specific location at the side of the vehicle and it is given as a step function with a constant force of 3382 KN (760,000 lb) for a total duration of .002 s.

### 4. FINITE ELEMENT MODEL DESCRIPTION

Prior to finite element model generation of the vehicle, a three-dimensional model of the hull was developed using PATRAN [13, 14]. The first step in model generation using PATRAN is the development of an initial level 1 model of the APC which includes grids, lines and patches. Grids describe points on the model and lines represent edges while patches describe model surfaces. Hyperpatches were used to represent solid portions of the three-dimensional model. Thus a simple geometric model of the entire hull of the M113A with access openings was generated upon which the finite element model would be based. A geometric description of the vehicle hull is given in Figure 2.

Surface normals were checked for each group of elements to ensure conformity and some rearranging of elements was found to be necessary. Since the transverse bending response was considered to dominate the overall response problem at short stand-off, shell elements rather than 3-D continuum elements were used to represent the model. Four-noded shell elements were selected to model the entire hull assembly. Each rectangle denotes one element in this figure. Plots of nodes and element numbers generated by PATRAN are omitted here because of overcrowding. However, location of impact load on the side wall as well as some critical node locations at the top and bottom surfaces corresponding to secondary systems inside the vehicle extrapolated to the hull model are indicated on this figure by circles as shown.

To assess the influence of large multiple cutouts on the overall response of the hull model, it was decided to incorporate simplified cutout shapes only on the top hull surface in the vicinity of the critical locations for the secondary equipment attachment points. The large ramp door at the rear end was largely ignored for the first order model. The cutouts were



approximated as square or rectangular shaped for both driver's and commander's hatch as well as crew access openings to allow ease of modelling and inclusion into the existing basic M113 hull model. Future refined models could be generated from scratch to accommodate exact shapes and location of openings through a complicated mapping procedure which was avoided during this study.

The finite element model generated by PATRAN consisted of an assembly of 311 elements with 322 nodes. Thickness of each shell element was specified to be equal to the wall thickness of the APC hull which remained constant. For stress computation each element was allowed to have a 2x2x2 integration scheme. Isometric views of the finite element model of the vehicle from a fixed vantage point are shown in Figure 3a and 3b. Figure 3a depicts the top, the front and the left hand side wall surfaces while in Figure 3b these surfaces were removed as indicated by the addition of intermittent lines to enable the viewer to have a clear view of the bottom, rear and far side walls of the simulated model.

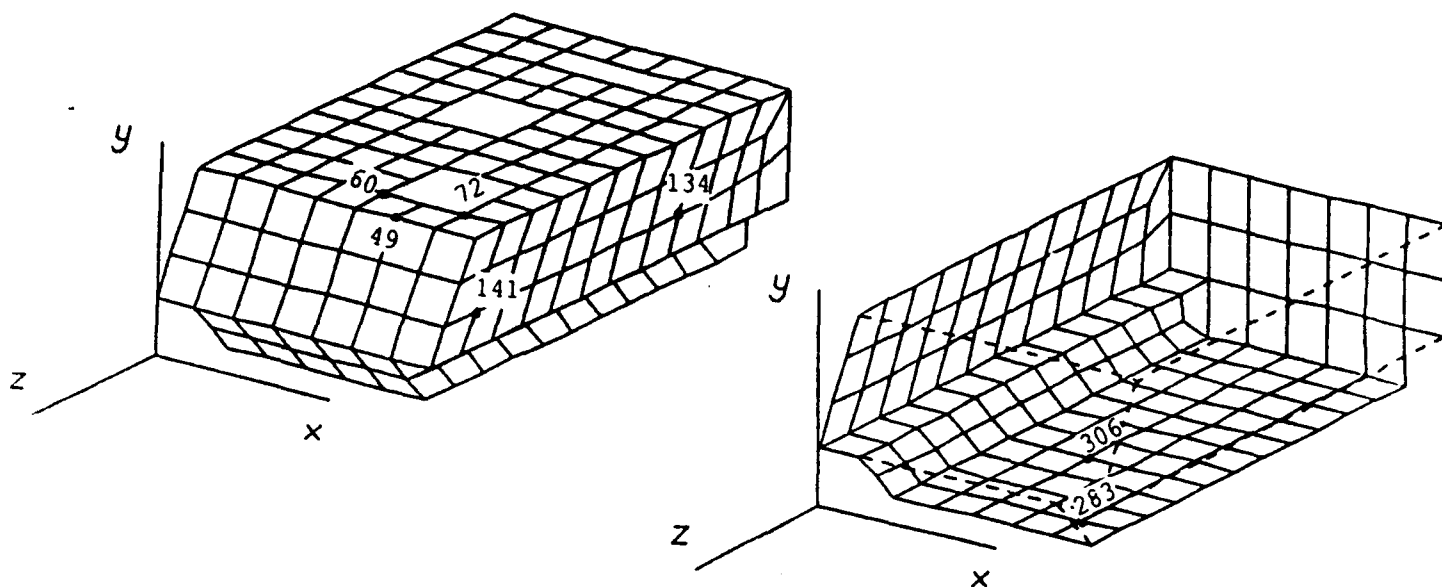
Initially the vehicle was not constrained from rigid body motions for displacement response predictions. However, this condition caused large rigid body displacements and rotations which dominated the dynamic response and exaggerated the distortional response. Similar problems were encountered during acceleration response computation. Large sliding motion of the vehicle for the unrestrained model would require correction by subtracting the rigid body displacement and rotation of the center of mass of the vehicle from the overall response. For subsequent studies a restrained model with 5 constrained degrees of freedom at three corner nodes on the bottom surface was used for computation of the transient response at critical locations of the vehicle since rigid body acceleration response could not be easily ascertained and eliminated from the overall response.

Due to a lack of ground dynamic friction coefficient data, the restraining effect of the ground friction upon the vehicle track and wheels at the contact zone could not be accurately modeled. However, the problem could be bracketed between two limiting conditions of unrestrained sliding and rotation corresponding to zero ground friction and fixed restraint corresponding to a peak ground friction allowing almost no sliding which could be achieved by fixing appropriate degrees of freedom at the bottom surface corner nodes of the vehicle. The limiting condition of fixed restraint has been represented in this current study.

## 5. MATERIAL MODEL

During the process of model generation material property identification numbers are assigned to the element groups as they are created. In the current investigation only one material identification was used to designate all groups of elements in the entire hull constructed from 5083 aluminum. Thickness of each shell element equalled the hull thickness of 3.175 cm which remained uniform throughout the structure.

For the transient response analysis, a linear elastic isotropic material model for the shell elements in the ADINA code was used. The finite element analysis employed the following values for 5083 aluminum: Young's modulus = 68,950 MPa. Poisson's ratio = 0.33, and mass density =  $2.7 \text{ g/cm}^3$ .



Node 134 - Center of Loading.

Node 141 - Instrument Panel Attachment Point.

Nodes 49, 62, 72 - Periscope Attachment Points.

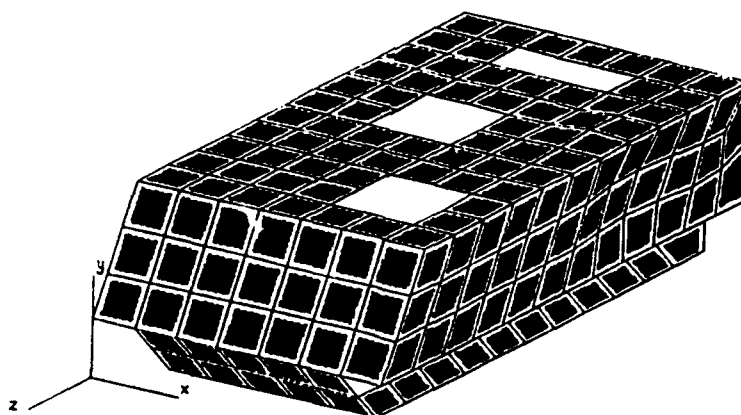
Figure 3a. Top and side wall node locations.

Node 283 - Driver's Seat Location.

Node 306 - Commander's Seat Location.

Figure 3b. Node locations at the bottom floor.

Figure 3. Critical node locations for the finite element model.



Time Step 22, Lumped Mass  
0.00220 Seconds

Deflection Magnified 1.00  
Scale 1/30, Alpha - 55.00, Beta - 71.50

Figure 4. Side wall deformation at .0022 s from impact.

## 6. DYNAMIC RESPONSE ANALYSIS

Prior to undertaking a dynamic response analysis of the structure subjected to transient loads, a free vibration analysis was conducted and the first 20 mode shapes and eigenfrequencies were determined using the lumped mass as well as the consistent mass formulations in the ADINA code. This study was useful for verification of the finite element model of the APC and indicated the type of deformation and the length of time it takes to respond to an unit impulse from an analysis of the mode shapes and time periods of oscillation of the structure.

The dynamic analysis focused on predicting structural deformation of the hull due to a side-on impact load. Additionally, location, magnitude and time of occurrence of peak displacements and accelerations at critical locations near sensitive equipments and crew positions are of interest due to ballistic shock propagation and lethal damage to coupled secondary systems. A nondestructive methodology for the assessment of vulnerability/survivability of armored vehicles and personnel carriers subjected to ballistic shock damage to the primary structure and propagation to the attachment points of coupled secondary sensitive systems due to nonpenetrating impact and close-in blast loads needs to be developed as a useful predictive tool from systematic reduction of time domain response to shock spectra at critical locations.

### 6.1 Results and discussions

The time domain response analysis was conducted using the Newmark implicit integration scheme and the lumped mass as well as the mode superposition method using the first 100 modes with the ADINA code in which we used a constant time step of .0001 s for the first 2000 cycles corresponding to a total response time of 0.2 s. Figure 4 shows the deformation response at .002 s corresponding to the occurrence of peak displacement at the impact point. The scale chosen for the hull configuration is 1/30 and the magnification factor used for deformation plot is 1.0. Considerable deformation of the vehicle hull can be observed at the sidewall in the vicinity of the impact point at node 134 where a concentrated time dependent step load of 3382 KN was applied initially and maintained at a constant level for a total duration time of .002 s.

For the model with lumped mass formulation, the maximum displacement at node 134 was computed to be 9.738 cm (3.834 in) which occurred at .0022 s. This compares favorably with the predicted peak resultant displacement of 9.726 cm (3.829 in) occurring at .002 s at the same location of the basic vehicle hull with no cutouts using the lumped mass formulation. The code predicted displacement and acceleration responses at specified node points in all three directions. However, the responses along the lateral X-direction were consistently at least an order of magnitude higher than those along the longitudinal and vertical directions of the vehicle. This is caused by the side-on impact load which is predominantly in the transverse thickness direction acting normally upon the sidewall. The displacement and acceleration responses of the left sidewall at node 134 along the X-coordinate direction obtained using the lumped mass formulation are shown in Figure 5. Peak displacement and acceleration responses occur at early times followed by lateral elastic

# X COMPONENT AT NODE 134

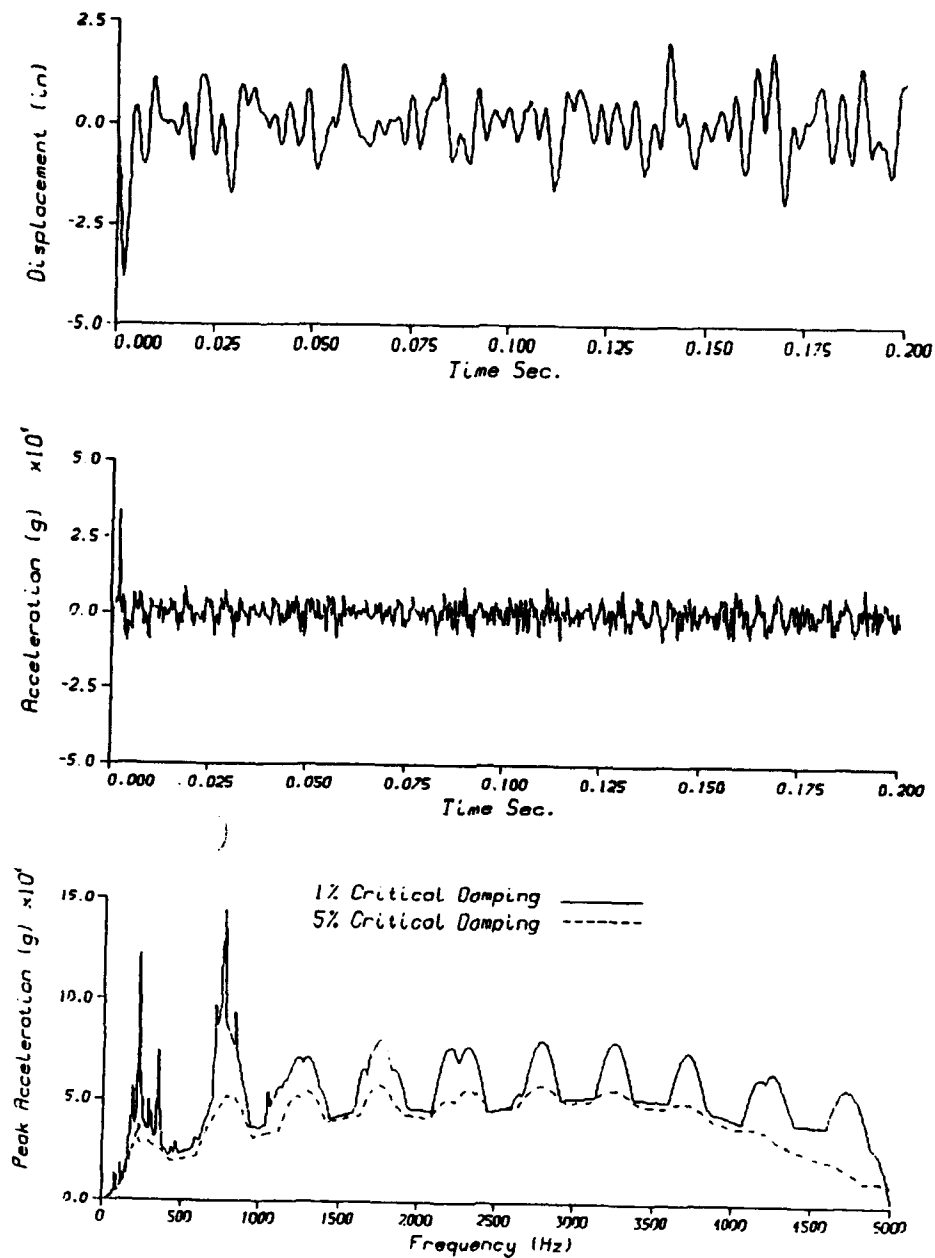


Figure 5. Displacement and acceleration responses at node 134 along the X-coordinate using the lumped mass formulation only.

oscillations.

The acceleration time domain data was analyzed and converted to a frequency domain shock response spectrum using a recursive filtering procedure [9]. A time step of 0.0001 s was selected for this computation. The dashed curve corresponds to a critical damping factor of 0.05 while the continuous curve employs a critical damping factor of 0.01. The influence of damping on the response spectra is clearly demonstrated in this figure. The shock spectrum analysis from the modal superposition method using first 100 modes indicates that the peak acceleration response is contained within the initial frequency range from 0-500 Hz followed by an uniformly spaced sequence of harmonics continuing at higher frequency levels. Results of similar computation using modal superposition and the first hundred vibratory modes exhibit similar trends as shown in Figure 6. Displacement and acceleration responses as well as an analysis of the shock spectra along the X-coordinate at node location 141 which is the site of the attachment point for the driver instrumentation panel to the left sidewall obtained using the mode superposition method in ADINA are shown in Figure 7. In this case the peak acceleration response appears to be concentrated in the low frequency range of 0-500 Hz. Computations using both mode superposition and lumped mass formulations at several other critical locations remote from the impact point have been completed. However, the predictions could not be included here due to a lack of space and will be presented later.

Comparison of shock spectra at critical locations remote from the cutout regions between the hull model with access openings and the basic hull without cutouts show very little change in shock response characteristics since shock propagation from the impact point to these remote locations such as the driver's and the commander's seat locations projected to the bottom floor remains relatively unaffected. However, significant change in peak acceleration magnitudes and shock response characteristics could be observed in all three coordinate directions at the three periscope locations in the vicinity of the simulated cutout for the driver's hatch opening and also to some extent at the instrument control panel attachment point at the front left hand side wall. These are caused possibly by the alteration of shock propagation path to the critical locations due to the presence of cutouts in the immediate neighborhood contributing to reduction in mass and stiffness which also affect the transient response of the vehicle at these locations.

## 6.2 Conclusions

Modern combat vehicles are increasingly carrying a variety of sensitive components which are susceptible to shock damage. The transmission of shock through the structure to critical components considerably remote from the impacted zone can result in the loss of combat capability inspite of crew survivability and retention of structural integrity of the vehicle.

As a first step towards the development of a vulnerability assessment methodology to predict damage to critical components from transient loads, a simplified model (1738 degrees-of-freedom) was generated for use in the ADINA code to compute the response of the U.S. Army M113 personnel carrier. A shock response analysis was applied to the acceleration histories to obtain shock spectra at these locations. The spectra at critical locations indicated that the simplified model contained no structural frequencies above 5

# X COMPONENT AT NODE 134

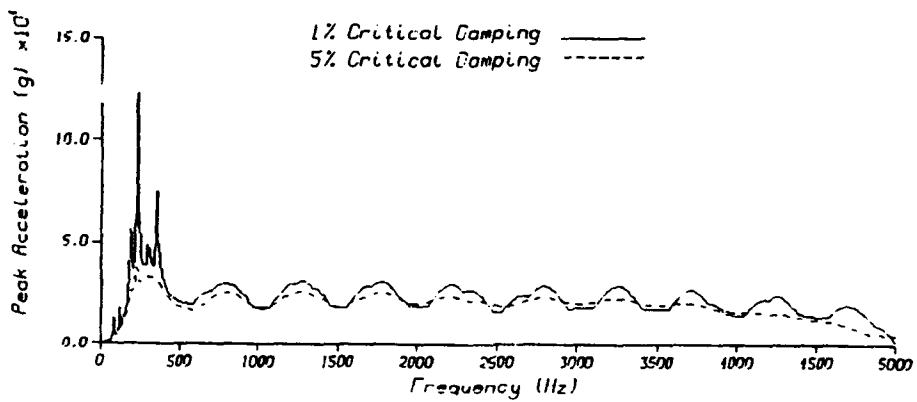
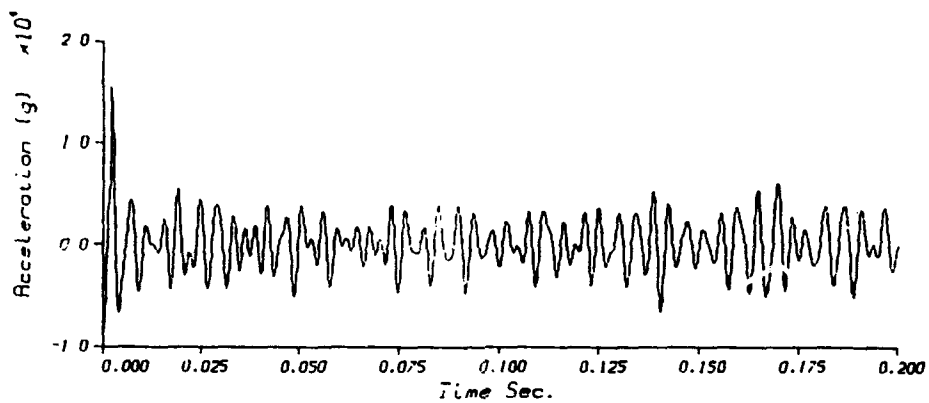
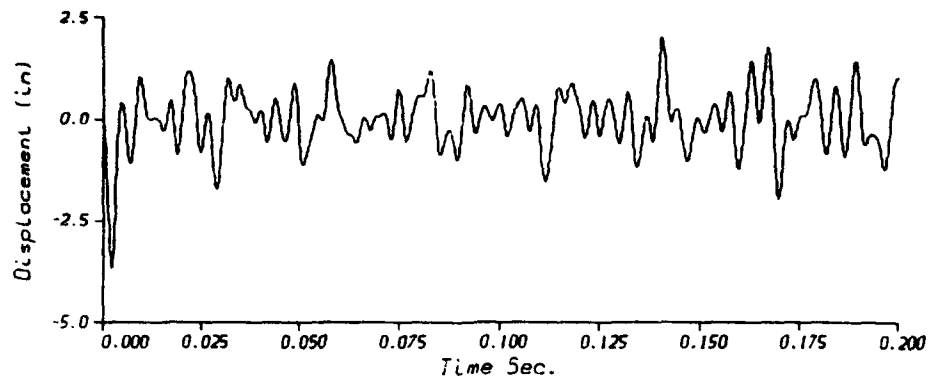


Figure 6. Displacement and acceleration responses at node 134 along the X-coordinate using mode superposition and lumped mass formulation.

# X COMPONENT AT NODE 141

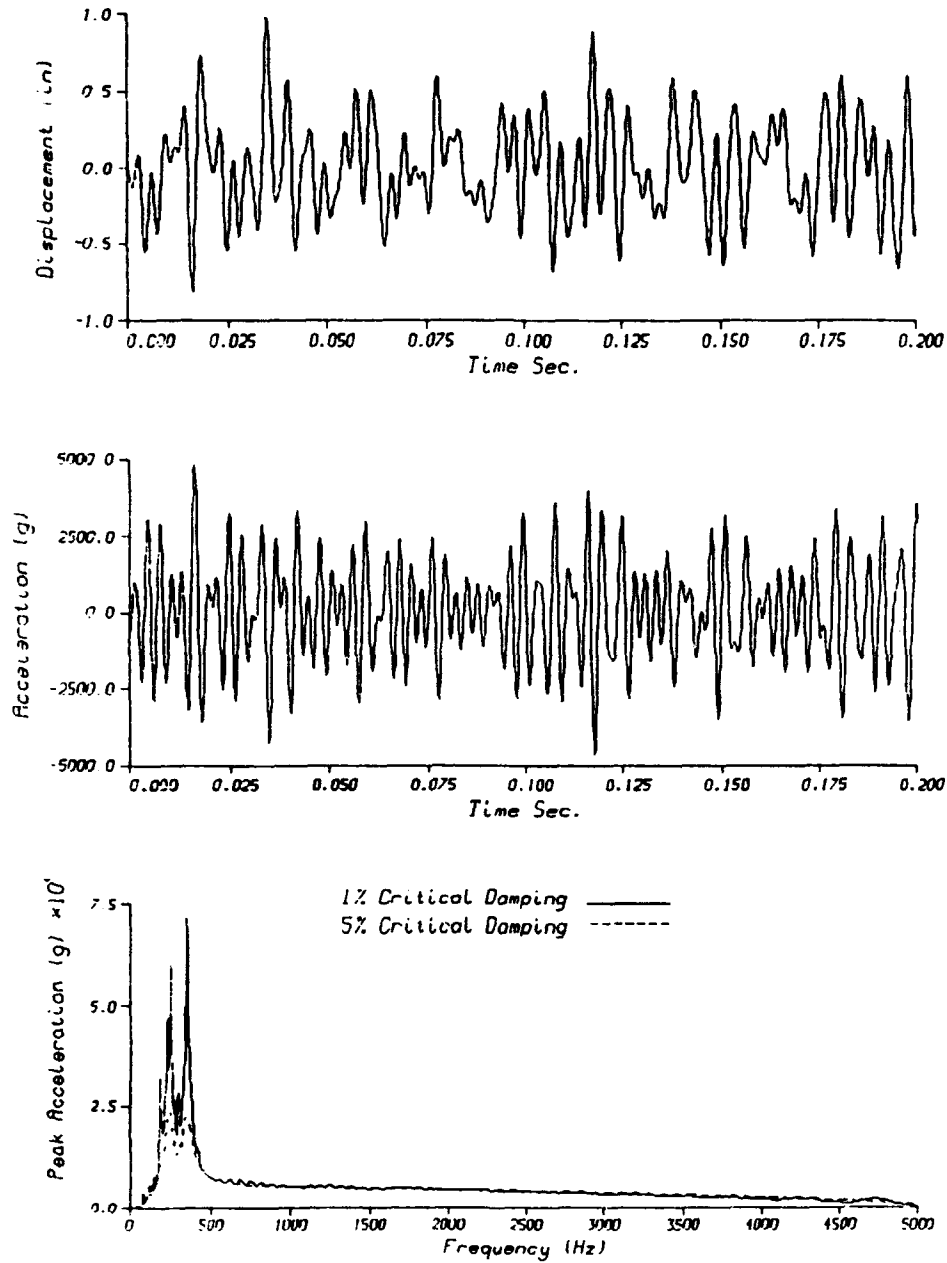


Figure 7. Displacement and acceleration responses at node 141 along the X-coordinate using mode superposition and lumped mass formulation.

KHz. Additionally, the spectrum at the impact point exhibited an uniformly spaced sequence of harmonics continuing beyond the cut-off frequency of 5 KHz and having periods equal to multiples of the load duration. Calculations were repeated using modal superposition based on the first 100 vibratory modes with a cut-off frequency of approximately 500 Hz. Again no structural frequencies higher than the 500 Hz cut-off were detected, but the same sequence of harmonics were observed at the impact point. These harmonics are an expected artifact of the load duration and has time periods which are multiples of the the impact load duration time of 0.002 s. The modal superposition calculation demonstrated that while the deflection histories can be accurately reproduced with the first 100 modes, the acceleration histories and shock spectra may require considerably higher number of modes. Some problems with aliasing resulting from discrete sampling of time signals were encountered but could be avoided by increasing the sampling rate. The study indicates that the time integration and superposition methods in combination with the shock spectral techniques can be useful in predicting damage due to shock propagation in structures.

Comparison of response of the model with access openings with that of the basic vehicle hull with no cutouts indicates a small shift of the eigenfrequencies to somewhat higher natural frequencies with very minor alteration in corresponding modeshapes for the cutout model possibly due to the mass loss in the cutout regions. Both frequency domain and time domain response at critical locations remote from the cutout regions for the vehicle model with access holes compare favourably with the response of the basic vehicle since propagation of shock to these regions remain unaffected. However, in the vicinity of the cutouts, shock propagation to critical regions are significantly affected due to alteration of the shock propagation path and reflection of energy from the interface of the cutout back to the source resulting in corresponding influence on the shock spectral response and visible alteration in peak frequency magnitudes when compared to those from the basic vehicle hull without any access openings.

## REFERENCES

1. Henried, A.G., and Jeng, J.M., "Dynamic Response of Secondary Systems in Structures Subjected to Ground Shock or Impact," Eng. Struct., Vol. 9, pp. 19-26, 1987.
2. Kelly, J.M., and Sackman, J.L., "Shock Spectra Design Methods for Equipment-Structure Systems," The Shock and Vibration Bulletin, No. 49, Part 2, pp. 171-176, 1979.
3. Nicholson, J.W., and Bergman, A., "Vibration of Damped Plate-Oscillator Systems," Journal of Engineering Mechanics, Vol. 112, No. 1, pp. 14-30, 1986.
4. Gupta, A.D., Wisniewski, H.L., and Bitting, R.L., "Response of a Generic Vehicle Floor Model to Triangular Overpressure Loads," Computers and Structures, Vol. 32, No. 3/4, pp. 527-536, 1989.
5. Walton, W. Scott, "New Ballistic Shock Protection Requirement for Armored Combat Vehicles," Proceedings of the 60th Shock and Vibration Symposium, Vol. 1, Virginia Beach, VA, November 1989.



6. Walton, W. Scott, "Propagation of Ballistic Shock in a Flat Steel Plate," Proceedings of the 61st Shock and Vibration, Vol. II, Pasadena, CA, October 1990.
7. Quigley, E.F., "EPIC-2 Predicted Shock Environments for Non-perforating Ballistic Impact," BRL-TR-2886, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, January 1988.
8. Quigley, E.F., "EPIC-2 Calculated Impact Loading History for Finite Element Analysis of Ballistic Shock," BRL-TR-3058, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, December 1989.
9. Kelly, R., and Richman, G., "Principles and Techniques of Shock Data Analysis", Report No. SVM-5, pp 142-146, Shock and Vibration Information Center, Naval Research Laboratory, Washington, DC, 1969.
10. Foss, C.F., "Jane's World Armoured Fighting Vehicles," St. Martin's Press, Inc., 175 Fifth Avenue, New York, N.Y., pp. 294-296, 1976.
11. Private Communication with Program Manager's Office, Light Combat USATACOM, Warren, MI, May 1991.
12. "ADINA - a Finite Element Program for Automatic Dynamic Incremental Nonlinear Analysis," Report ARD 90-1, ADINA R and D Inc., Watertown, MA, 1990.
13. PATRAN User's Guide, PDA Engineering Software Products Division, Santa Ana, CA, 1990.
14. PATRAN/ADINA Interface Guide, Version 4, PDA Engineering PATRAN Division, Santa Ana, CA, 1990.

## **The Scaling Laws for Fluid Mixing**

*Qiang Zhang and James Glimm*

Department of Applied Mathematics and Statistics  
SUNY at Stony Brook  
Stony Brook, NY 11794-3600

### **ABSTRACT**

We have developed a quantitative theory for the mixing of fluids induced by a random velocity field. The theory provides a quantitative prediction for the growth of the mixing region. There are three distinct regimes for the asymptotic scaling behavior of the mixing layer, depending on the asymptotic behavior of the random velocity field. The asymptotic diffusion is Fickian when the correlation function of the random field decays rapidly at large length scales. Otherwise the asymptotic diffusion is non-Fickian. The scaling behavior of the mixing layer driven by a general random velocity field is determined over all length scales. Our results show that, in general, the scaling exponent of the mixing layer is non-Fickian on all finite length scales. In the Lagrangian picture, due to the non-linearity of the effective dynamical equation derived from the Taylor diffusion theory, the mixing layer is not a fractal even if the random velocity field is a fractal.

### **Introduction**

The study of the mixing induced by random fields becomes increasingly important in the study of fully developed turbulence, enhanced oil recovery processes and ground water ecology. For a tracer flow through a random velocity field, a mixing layer is developed between the tagged and the untagged region. The mixing region expands as time evolves. In the case of ground water ecology, the random velocity field is caused by the random permeability field through the Darcy's law. The purposes of our study is to predict the statistical properties of the fluids, such as the size of the mixing regime, or the

effective macroscopic diffusion coefficient, from the statistical properties of the random field. Let  $l(t)$  be the size of the mixing at the time  $t$ . How does  $l$  grow as a function of time  $t$ ? At what rate will  $l$  grow at very large time? In general,  $l$  can take a quite general functional form. Let us first define the scaling exponent of a general function. For any positive differentiable function  $f(t)$ , we express  $f(t)$  in a multi-length-scale-fractal form

$$f(t) = a(t)t^{b(t)}, \quad (1)$$

where

$$b(t) = \frac{d \ln f(t)}{d \ln t} \quad (2)$$

is the multi-length-scale fractal exponent at scale  $t$  and  $a(t)$ , determined by

$$\ln a(t) = \ln f(t) - \ln(t) \frac{d \ln f(t)}{d \ln(t)}, \quad (3)$$

is the multi-fractal coefficient of  $f(t)$ . In general, the exponent of a function may vary with the scale. If  $b(t)$  is constant, then it follows that  $a(t)$  is also a constant. The result is a pure power law. In this case,  $f(t)$  is a fractal in the sense that results obtained from different scales are related by a simple scaling. Knowing the value of  $f$  at a given length scale and its scaling exponent  $b$ , the value of  $f$  over all length scales can be determined by a simple scaling. However, the fractal treatment of the mixing layer is for simplicity, rather than necessity. Even if we take the permeability field as a fractal random field, the velocity field obtained from Darcy's may not be a fractal field, except for a weak heterogeneity where the linear (in terms of the fluctuation of the permeability field) solution of Darcy's law can be used. In general, the statistical properties of a random field can vary as the length scale changes. Therefore a multi-fractal theory is more suitable for studies involving a wide range of length scales, such as data including both laboratory experiments and field tests [21]. The length scale in the laboratory is on the order of a fraction of a meter and the length scale of the interwell spacing in an oil field may be on the order of a kilometer.

By applying the renormalization group perturbation theory, we have developed a multi-length-scale-fractal theory for the mixing of the fluids induced by random fields [14,24,25,28]. The theory provides an explicit analytic prediction for the growth of the mixing regime of the fluids and an effective macroscopic diffusion coefficient induced by random velocity fields, or the heterogeneous permeability fields, over all time scales.

The results of the exact numerical computation using the front tracking method agree with the prediction of the theory very well over all time scales without any fitting parameters [7-11].

The study of diffusion induced by random field has a long history. Many methods have been developed. See [1-6,12,13,15-20,22,23] for further references.

### Asymptotic Scaling Relation

The relationships among the asymptotic exponents of the fluid, diffusion coefficient and random field have been by several different methods [14,24,25,28]. These methods gave the same results. The results are

$$\gamma_{\infty} = \max\left\{\frac{1}{2}, 1 + \frac{\beta_{\infty}}{2}, 1 + \frac{\alpha\beta_{\infty}}{2}\right\} \quad (4)$$

and

$$\psi_{\infty} = \max\{0, 1 + \beta_{\infty}, 1 + \alpha\beta_{\infty}\}. \quad (5)$$

Here  $\gamma_{\infty}$ ,  $\psi_{\infty}$  and  $\beta_{\infty}$  are the asymptotic scaling exponents of the mixing regime, of the effective macroscopic diffusion coefficient, and of the correlation function of the velocity (or permeability) field respectively.  $\alpha$  is a parameter that characterizes the rate of removal of the infrared cut-off. Equations (4) and (5) show that there are three regimes with distinct scaling behavior. These three regimes are connected by two critical points:  $\beta_{\infty} = -1$  and  $\beta_{\infty} = 0$ . For an asymptotically rapidly decaying velocity correlation function,  $\beta_{\infty} < -1$ , the asymptotic diffusion is Fickian. This regime is characterized by the properties that the correlation function is integrable over one spatial dimension and that a finite asymptotic limit exists for the diffusion coefficient. For asymptotically slowly decaying velocity correlation function,  $-1 < \beta_{\infty} < 0$ , the asymptotic scaling is non-Fickian and the diffusion coefficient diverges at large time. The analysis of the field data shows that this regime is important for ground water ecology. The third regime is specified by  $0 < \beta$ . In this regime, the scaling is non-Fickian. A striking new result has been obtained in this regime: the asymptotic scaling is non-unique before specifying how the infra-red cutoff is removed from the system. [25,28]. In other words the scaling relation contains a family of solutions which depends on the rate of removal of infrared cut-off. Such non-uniqueness is intrinsically due to the fact that infrared cutoff of the correlation function of the random velocity field is not removable for  $0 < \beta_{\infty}$ . As  $\alpha$  tends to

infinity, which corresponds to the fastest removal of the infrared cutoff, the scaling exponent tends to infinity as well.

For a laminar shear flow model with a fractal Gaussian random velocity field, the problem can be treated exactly [3,28]. It has been shown that the solution given in [3] corresponds to a particular solution with a specific infrared cutoff among the family of solutions given in [28]. In the case of an infrared non-removable system, any specific choice for the infrared cutoff may give spurious phase boundaries in the asymptotic scaling laws.

### Transient Scaling Behavior

The scaling laws given by (4) and (5) are valid for very large scales only. We now consider the scaling laws on finite length scales. In the Eulerian picture, the motion of the tracer flow can be modeled by a stochastic linear transport equation

$$s_t + \vec{v} \cdot \nabla s = d_l \nabla^2 s, \quad (6)$$

where  $s$  is the saturation value of the fluid, so that  $s = 1$  for tagged fluid and  $s = 0$  for untagged fluid.  $\nabla$  is the spatial gradient operation.  $d_l$  is the local molecular diffusion coefficient.  $\vec{v}$  is the velocity of the fluid. We assume the velocity field is random and stationary. Therefore it is a function of the spatial variables only. For tracer flow, the velocity field  $\vec{v}$  is determined from Darcy's law and the condition of incompressibility. Each individual solution  $s(t, \vec{x}, \vec{v})$  for a given realization of the random variable  $\vec{v}$  does not play a significant role. The statistical behavior of the tracer is determined from the ensemble mean saturation value  $\langle s(t, \vec{x}) \rangle$ . Although (6) is a linear equation with respect to the independent variables  $t$  and  $\vec{x}$ , it is a non-linear equation respect to the random fields  $\vec{v}$  and  $s$ . The effective equation for  $\langle s \rangle$  will not be obtained from (5) in a trivial way. It has been shown that the effective equation for  $\langle s \rangle$  can be written as [14, 24, 25]

$$\begin{aligned} \frac{\partial \langle s(t, \vec{x}) \rangle}{\partial t} + \vec{v}_0 \cdot \nabla \langle s(t, \vec{x}) \rangle &= \nabla \cdot \int_0^t \langle \delta \vec{v}(\vec{v}_0 t) \delta \vec{v}(\vec{v}_0 t') \rangle dt' \cdot \nabla \langle s(\vec{\eta}, t) \rangle \\ &+ d_l \nabla^2 \langle s(\vec{\eta}, t) \rangle + O(\delta \vec{v}^4). \end{aligned} \quad (7)$$

Here  $\vec{v}_0 = \langle \vec{v} \rangle$  is a constant vector and is determined by the steady pressure drop applied to the fluid field.  $\langle \cdot \rangle$  denotes ensemble average.  $\delta \vec{v} = \vec{v} - \vec{v}_0$  is the fluctuation of the

random velocity field due to the rock heterogeneity and is a function of the spatial variables only for a stationary velocity field. The shape of the tracer interface at  $t = 0$  is given by  $s(0, \vec{x})$ . For simplicity we assume that the initial interface between tagged and untagged fluid is a plane with its normal direction pointing along the direction of  $\vec{v}_0$ . Then the effective equation in the longitudinal direction is given by [14, 24, 27]

$$\frac{\partial \langle s(t, x) \rangle}{\partial t} + \vec{v}_0 \partial \langle s(t, x) \rangle = c - \frac{\partial^2}{2} \langle s(t, \vec{x}) \rangle + O(\delta v^4). \quad (8)$$

Here

$$\alpha(t) = \int_0^t \langle \delta v_1(v_0 t) \delta v_1(v_0 t') \rangle + d_l = \int_0^t q(v_0(t - t')) + d_l \quad (9)$$

is the effective longitudinal diffusion coefficient.  $\delta v_1$  is the longitudinal component of the fluctuation of the random velocity field.  $q$  is the longitudinal velocity correlation function. In (19) we have assumed that the statistical properties of the random velocity are translational invariant. Therefore the correlation function of the velocity field depends only on the relative separation between the two points. In the limit of weak fluctuations, the term proportional to  $\delta v^4$  is negligible. Then for the planar initial data, the solution to (8) is given by

$$\langle s(x, t) \rangle = \frac{1}{2} \text{erfc}\left(\frac{x - v_0 t}{l(t)}\right). \quad (10)$$

where  $\text{erfc}()$  the complimentary error function and

$$l(t) = 2\left[\int_0^t \alpha(t') dt' + d_l t\right]^{1/2} = 2\left[\int_0^t (t - t') \alpha(t') dt' + d_l t\right]^{1/2}. \quad (11)$$

Equation (10) shows that the solution  $\langle s(x, t) \rangle$  scales as  $l(t)$ . It follows that the mixing zone scales as  $l(t)$  as well. We define  $l(t)$  as the mixing length, the size of the mixing zone. From definition (2), scaling exponent of the mixing regime is given by

$$\gamma(t) = \frac{d(\ln(l(t)))}{d(\ln(t))} = \frac{1}{2} \left[ 1 - \frac{\int_0^t \xi q_i(\xi) d\xi}{t \int_0^t q(\xi) d\xi + d_l t} \right]^{-1}, \quad (12)$$

and the scaling exponent of the effective diffusion coefficient is given by

$$\psi(t) = \frac{d(\ln(\alpha(t)))}{d(\ln(t))} = \frac{1}{2} \left[ 1 - \frac{\int_0^t \xi q_i(\xi) d\xi}{t \int_0^t q(\xi) d\xi + d_l t} \right]^{-1}. \quad (13)$$

The first two regimes of the asymptotic scaling relations shown in (4) and (5), i. e.  $\max\{\frac{1}{2}, 1 + \frac{\beta_\infty}{2}\}$  and  $\max\{0, 1 + \beta_\infty\}$  can be determined from (12) and (13)[14,24]. It can be shown that, at the critical point  $\beta_\infty = -1$ , the system has the slowest rate for approaching its asymptotic exponent [24]. Equations (12) and (13) are only valid for the systems in which the infrared cutoff is removable. In the case where the infrared cutoff can not be completely removed, one must consider the velocity correlation function in Fourier space with an infrared cutoff. Then the asymptotic scaling exponents depend on the rate of the removal of the infrared cutoff and a family of solutions exist. See [28,25] for the analysis of asymptotic scaling exponent of the infrared non-removable system.

Equations (12) and (13) show that the exponent of the mixing length and that of the diffusion coefficient at length scale  $\nu_0 t$  depend on all length scales less than  $\nu_0 t$ . This explains the dependence of the mixing length exponent on the flow history. In general, the ratio

$$\frac{\int_0^t \xi q(\xi) d\xi}{t \int_0^t q(\xi) d\xi + d_l t}$$

is non-zero for finite  $t$  unless  $q$  is proportional to a delta function. It follows from (12) and (13) that, in general, the scaling of the mixing zone is non-Fickian or finite length scales unless the random velocity is a white noise.

Equations (4) and (5) show that molecular diffusion does not play a role in the asymptotic scaling exponents of the mixing length and diffusion coefficient. The molecular diffusion is important on small length scales only. Assuming  $q$  is non-singular on small length scales, then from (12) we have  $\gamma(0) = 1/2$  when  $d_l \neq 0$ , and  $\gamma(0) = 1$  when  $d_l = 0$ . For small length scales, (15) gives

$$\gamma(t) \approx \frac{q(0)t + d_l}{q(0)t + 2d_l}.$$

Therefore at the length scale  $\nu_0 t \gg \nu_0 d_l / q(0)$  the molecular diffusion is negligible and  $\gamma(t) \approx 1$ . The data set given in [21] shows that scale exponent  $\gamma$  is about 1 at the smallest length scale in the given data. This implies that the local molecular diffusion is negligible at all length scales for the given data set, including the short length scale data obtained from laboratory experiments. Equation (12) also shows that for a given

correlation function and length scale  $\nu_0 t$ ,  $\gamma$  is a decreasing function of the strength of the molecular diffusion. In other words, for fixed  $t$ , the larger  $d_l$  is, the smaller  $\gamma$  is. This is due to the fact that the constant local molecular diffusion has Fickian scaling while the diffusion induced by a random velocity has non-Fickian scaling (except when the random velocity is a white noise).

The properties discussed here are in good agreement with the general features of the laboratory and field data [21]. This data set shows that the exponent of the mixing layer is about 1 at short length scales, and is less than 1 at large length scales. It also confirms that the exponent of the mixing region is larger than 1/2, i.e. non-Fickian over all length scales supplied by the data set.

The main features of the multi-fractal theories relative to the fractal theories, are the allowance of a length scale dependence of physical quantities and the inclusion of the transient effect between different length scales. Equations (12) and (13) show that, at any given length scale  $\nu_0 t$ , the scaling exponent of the mixing layer and that of the diffusion coefficient depend on the behavior of the correlation function of the random velocity on all length scales which are smaller than  $l$ . This is a transient effect. Our analysis has shown that two factors contribute to the variation of the scaling exponent of the fluid at different length scales. One is the variation of the scaling exponent of the random velocity field at current length scale. The other is the variation of the scaling exponent of the random velocity field at all length scales smaller than the current length scale (a transient effect). If the exponent of the velocity correlation function  $f$  varies slowly over a sufficient large range of length scales, then the exponent of the mixing layer,  $\gamma(t)$ , approaches an instantaneous fractal exponent  $\gamma_{fr}(t) = \max\{\frac{1}{2}, 1 + \frac{\beta(t)}{2}\}$ , where  $\beta(t) = d \ln q(t) / d \ln t$ . In the asymptotic regime, the scaling exponent of the velocity correlation function approaches its asymptotic limit, and the transient effect from finite length scales is damped. The asymptotic scaling exponent of the fluid is determined by the asymptotic scaling exponent of the random velocity field only. Therefore fractal random velocity field models can be used for the study of the asymptotic scaling relationship between the fluid and the random velocity field [14,25]. For a random velocity field with an asymptotic scaling exponent less than  $-1$ , the asymptotic scaling behavior of the fluid is Fickian. In this case, a pure fractal random velocity field leads to an ultraviolet divergence. Any cutoff, which removes such divergence, changes the random velocity field



from a fractal field to a multi-fractal field. Since the introduction of the ultraviolet cutoff only affects the scaling behavior of the random velocity field at short length scales, it does not affect the asymptotic scaling relation between the fluid and the random velocity field. The situation is quite different for a random velocity field with an asymptotic scaling exponent larger than zero. In this case a pure fractal leads to an infrared divergence. Since the infrared cutoff changes scaling behavior of the random velocity field at large length scale, it intrinsically affects the asymptotic scaling relationship between the fluid and velocity field. Therefore, for infrared divergent systems, the asymptotic scaling relationship between the fluid and the random velocity field depends on the rate at which the infrared cutoff tends to zero as shown in (4) and (5) [28,25].

The multi-length-scale-fractal theory presented here has been validated by the exact numerical computations using the front tracking method [7-10]. The comparisons between the exact numerical results and the results predicted by the multi-length-scale-fractal theory have been made for a system with asymptotic non-Fickian diffusion ( $\beta_\infty = -0.5$ ) and a system with asymptotic Fickian diffusion ( $\beta_\infty = -\infty$ ). Different strengths of heterogeneity are studied for each system. The studies show that the predictions of multi-length-scale-fractal theory agree very well with the numerical results over the full range of the length scales computed which is up to the length scale 100 times larger than the length scale of the basic heterogeneity. We comment that in these studies, no fitting parameter is involved for the results of the multi-length-scale-fractal theory.

According to (4), the asymptotic scaling exponent of the mixing layer is 0.75 for  $\beta_\infty = -0.5$ . The scaling exponent  $\gamma(t)$  obtained from the numerical computations decreases monotonically from its initial value close to 1.0 to its final value 0.85 at the end of the computation. In other words at the length which is 100 times larger than the length scale of the basic heterogeneity, the scaling exponent still has not reached its asymptotic limit. For  $\beta_\infty = -\infty$ ,  $\gamma_\infty = 0.5$ . That limit is approached within 10% at a length scale larger than 50 units of the basic heterogeneity length scale. Both these two cases show that the transient effects are very important for determination of the scaling on finite length scales. See [7-10] for the details of these numerical studies.

The study of the diffusion induced by random velocity field has also been carried out in the Lagrangian picture by applying Taylor diffusion theory [5, 6, 22, 25, 26]. Under Corrsin's hypothesis, the effective equation for the variance tensor of fluctuations

of the tracer particle displacement is a complicated non-linear ordinary differential equation. Recently, it has been shown that due to the nonlinearity of the effective equation, the mixing regime induced by a random field is not a fractal, even if the random field itself is a fractal [25,26]. In addition to the asymptotic scaling exponents, the asymptotic coefficients in front of the asymptotic power law in front of the asymptotic power law for the size of mixing layer and effective diffusion coefficient have also been determined for a fractal random velocity field. See [25] and [26] for details.

### Acknowledgement

The research of Q. Zhang is supported in part by the U. S. Department of Energy, contract DE-FG02-90ER25084 and the research of J. Glimm is supported in part by the U. S. Department of Energy, contract DE-FG02-90ER25084, the U. S. Department of Energy, contract DE-FG02-90ER25084, the National Science Foundation, grant DMS-8901884, and the Army Research Office, grant DAAL03-K-0017.

### REFERENCE

1. Y. Amirat, K. Hamdache and A. Ziani, *Homogénéisation d'équations hyperboliques du premier ordre et application aux écoulements miscibles en milieu poreux*, *Ann. Inst. Henri Poincaré*, 6 397 (1989).
2. A. Arya, T. A. Hewett, R. G. Larson and L. W. Lake, *Dispersion and reservoir heterogeneity*, SPE 15386, 60th Ann. Tech. Conf. of SPE, Las Vegas, Sept. 22.30, 1985.
3. M. Avellaneda and A. Majda, *Mathematical Models with Exact Renormalization for Turbulent Transport*. *Commun. Math. Phys.* 131, 381 (1990).
4. J. H. Cushman, *Development of stochastic partial differential equations for subsurface hydrology*, *Stochastic Hydrol. Hydraul.* 1 241 (1987).
5. G. Dagan, *Solute transport in heterogeneous porous formations*, *J. Fluid Mech.* 145 151 (1984).
6. G. Dagan, *Flow and Transport in Porous Formations*: Springer-Verlag, New York, 1989.

7. F. Furtado, J. Glimm, B. Lindquist and F. Pereira, *Multi-length scale computations of the mixing length growth in tracer flow*, proceedings of the Emerging Technologies Conference, Kovarik F. ed. 251, Houston TX, July 1990.
8. F. Furtado, J. Glimm, B. Lindquist and F. Pereira, *Characterization of mixing length growth for flow in heterogeneous porous media*, SPE 21283, Proceedings of the 11th SPE Symposium on Reservoir Simulation, Feb. 1991.
9. F. Furtado, J. Glimm, B. Lindquist, F. Pereira and Q. Zhang, *Time Dependent Anomalous Diffusion for Flow in Multi-fractal Porous Media*, Proceedings of the Workshop on Numerical Methods for the Simulation of Multiphase and Complex Flow. Lecture Notes in Physics, Verheggan T. M. ed., Springer Verlag, New York, (1991).
10. J. Glimm, W. B. Lindquist, F. Pereira and R. Peierls. *The Multi-fractal Hypothesis and Anomalous Diffusion*, Revista Brasileira de Matemamatica Aplicada e Computacional, 1992.
11. J. Glimm, W. B. Lindquist, F. Pereira and Q. Zhang. *A Theory of Macrodispersion for the Scale Up Problem, Transport in Porous Media*, to appear.
12. L. W. Gelhar, *Stochastic subsurface hydrology from theory to applications*. *Water Resources Res.* **22** 153S (1986).
13. L. W. Gelhar and C. L. Axness, *Water Resources Res.* **19** 161 (1983).
14. J. Glimm and D. H. Sharp, *A random field model for anomalous diffusion in heterogeneous porous media*, *J. Stat. Phys.* **62** 415 (1991).
15. A. L. Gutjahr and L. W. Gelhar, *Stochastic Models of Subsurface Flow: Infinite Versus Finite Domains and Stationarity*, *Water Resources Res.* **17** 337 (1981).
16. T. A. Hewett, and R. A. Behrens, *Conditional Simulation of Reservoir Heterogeneity with fractals*, SPE 18326. SPE Annual Technical Conference and Exhibition, Houston Oct. 2-5, (1988).
17. H. Holden and N. H. Risebro, *Stochastic Properties of the Scalar Buckley-Leverett Equation*, *SIAM J. Appl. Math.* **51** 1472 (1991)
18. R. H. Kraichnan, *Eulerian and Lagrangian renormalization in turbulence theory*, *J. Fluid Mech.*, **83** 349 (1977).

19. L. Lake and H. Carroll, eds., *Reservoir Characterization*, Academic Press, 1986.
20. W. D. McComb, *The Physics of Turbulence*, Oxford.: Oxford University Press, 1990.
21. S. P. Neuman, *Universal Scaling of Hydraulic Conductivities and Dispersivities in Geologic Media*, *Water Resources Res.* **26** 1749 (1990).
22. S. P. Neuman and Y. K. Zhang, *A Quasi-Linear Theory of Non-Fickian Subsurface Dispersion*, *Water Resources Res.* **26** 887 (1990).
23. L. Tartar, *Nonlocal effects induced by homogenization*, in *Partial Differential Equations and the Calculus of Variations*, Vol. 2, Essays in Honor of Ennio De Giorgi, 925-938. Birkhauser, (1989).
24. Q. Zhang, *A Multi-length Scale Theory of the Anomalous Mixing Length Growth For Tracer Flow In Heterogeneous Porous Media*, *J. Stat. Phys.* **66** 485 (1992).
25. Q. Zhang, *The Asymptotic Scaling Behavior of Mixing Induced by a Random Velocity Field*, *Advances in Applied Mathematics*, to appear.
26. Q. Zhang, *The Transient Behavior of Mixing Induced by a Random Velocity Field*, Preprint SUNYSB-AMS-92-08.
27. Q. Zhang, *Length Scales, Multi-Fractal and Non-Fickian Diffusion*, Proceedings of the IX International Conference on Computational Methods in Water Resources, June 9-12, 1992, Denver CO. vol. 2: Mathematical Modeling in Water Resources, pp. 59-70, (1992). Eds. T.F.Russell, R.E.Ewing, C.A.Brebbia, W.G.Gray and G.F.Pinder.
28. Q. Zhang and J. Glimm, *Inertial Range Scaling of Laminar Shear Flow as a Model of Turbulent Transport*. *Comm. in Math. Phys.* **146** 217 (1992).

# Stability Analysis of Stochastic PDE's via Lyapunov Functionals<sup>1</sup>

Pao-Liu Chow

Department of Mathematics

Wayne State University

Detroit, Michigan 48202

**ABSTRACT** The paper is concerned with the stability of stochastic partial differential equation of parabolic Itô type. By the method of Lyapunov functionals, we examine three kinds of stochastic stability: the stability in probability, the asymptotic moment stability and the almost sure asymptotic stability. Sufficient stability conditions are given and some illustrative examples are provided.

## I. INTRODUCTION

To illustrate some basic ideas in stochastic stability, we will first consider a simple example in stochastic ordinary differential equations (ODE's). The example is given by the following Itô equation in one dimension:

$$dx_t = -\alpha x_t dt + \sigma x_t db_t, \quad (1)$$

where  $\alpha, \sigma$  are positive parameters and  $b_t$  is the standard Brownian motion in one dimension. Clearly  $x_t \equiv 0$  is a solution. The question is whether the null solution is stable. To find it out, let

$$\varphi_t(x) = x^{2p}, \quad p > 0.$$

---

<sup>1</sup>This work was supported by the NSF grant DMS-91-01360.

By the Itô formula [1], we have

$$\begin{aligned}\varphi_p(x_t) &= \varphi_p(x_0) + 2p\left[\frac{1}{2}(2p-1)\sigma^2 - \alpha\right] \int_0^t \varphi_p(x_s) ds \\ &\quad + 2p\sigma \int_0^t \varphi_p(x_s) db_s,\end{aligned}$$

so that

$$E\varphi_p(x_t) = \varphi_p(x_0) + 2p\left[\frac{1}{2}(2p-1)\sigma^2 - \alpha\right] \int_0^t E\varphi_p(x_s) ds$$

or

$$E\varphi_p(x_t) = \varphi_p(x_0) \exp\left\{2p\left[\frac{1}{2}(2p-1)\sigma^2 - \alpha\right]t\right\}.$$

Thus, as  $t \rightarrow \infty$ , we get

$$E\varphi_p(x_t) \rightarrow \begin{cases} 0, & \text{if } \alpha > \frac{1}{2}(2p-1)\sigma^2, \\ \infty, & \text{if } \alpha < \frac{1}{2}(2p-1)\sigma^2, \end{cases}$$

which shows that the (asymptotic) moment stability depends on the order  $2p$ . For instance, for  $p = 1$ , the null solution is mean-square stable if  $\alpha = \frac{3}{4}\sigma^2$ , but is not 4th-moment ( $p = 2$ ) stable. In fact, for any given  $\alpha$  and  $\sigma$ , there exists a  $m > 0$  such that the null solution is  $m$ th-moment unstable. On the other hand, the equation (1) has the exact solution:

$$x_t = x_0 \exp\left\{\sigma b_t - \left(\alpha + \frac{1}{2}\sigma^2\right)t\right\}.$$

By the strong law of large numbers, we have  $(b_t/t) \rightarrow 0$  almost surely (a.s.) so that

$$x_t \rightarrow 0 \text{ a.s. as } t \rightarrow \infty,$$

for  $\alpha > -\frac{1}{2}\sigma^2$ . Hence the null solution is a.s. asymptotically stable for any  $\alpha \geq 0, \sigma > 0$ .

In contrast with a deterministic equation, this simple example illustrates many faces of a stochastic stability problem. For instance, the asymptotic stability can be discussed in

the sense of probability, in the  $p$ th-moment or in the almost sure sense. Which mode of convergence we choose depends on the nature of the problem under consideration.

Now consider a general Itô equation in one dimension:

$$dx_t = a(x_t)dt + \sigma(x_t)db_t, \quad (2)$$

where the functions  $a$  and  $\sigma$  are sufficiently smooth with  $a(0) = \sigma(0) = 0$ . Again  $x_t \equiv 0$  is a solution of Eq. (2). Since the equation is no longer solvable in a closed form, to determine the stability of the null solution, other approach is needed. As in the deterministic case, a qualitative method based on the Lyapunov function has been employed by many workers to study the stochastic stability. For a systematic exposition of this subject, one is referred to the excellent book by Khasminskii [2]. Since the emergence of stochastic partial differential equations (PDE's) in 1970's, the corresponding stability questions have arisen naturally. In 1982 the author introduced the method of Lyapunov functionals to study the stability of stochastic PDE's [3]. About the same time Ichikawa adopted a similar approach to analyze the stability and related questions [4]. Recently Khasminskii and Mandrekar [5] have examined the linearized stability problems of stochastic PDE's by the Lyapunov method.

In this paper we shall first introduce stochastic PDE's. Then we review some basic results in stochastic PDE's and define the Lyapunov functional. Subsequently, via the Lyapunov functional approach, we present some stability results in the sense of probability, the asymptotic stability in  $p$ th-moment and the almost sure asymptotic stability of some nonlinear stochastic PDE's. Some examples will also be given for the purpose of illustration.

## II. STOCHASTIC PDE'S

Let  $D = (0,1)$  be a unit interval and let  $K = L^2(D)$  be the Hilbert space of square-

integrable functions on  $D$ . For  $\varphi \in K$ , define the integral operator  $Q$  by

$$(Q\varphi)(x) = \int_D q(x, y) \varphi(y) dy, \quad (3)$$

where the kernel  $q(x, y)$  is symmetric and positive so that

$$\text{Tr}.Q = \int_D q(x, x) dx < \infty. \quad (4)$$

Let  $\{e_n\}$  be the set of orthonormal eigenfunctions of  $Q$  with eigenvalues  $\lambda_n$ ,  $n = 1, 2, \dots$ .

By (4), we have  $\lambda_n > 0$  and

$$\sum_{n=1}^{\infty} \lambda_n < \infty.$$

Denote by  $W(t, x)$  the  $Q$ -Wiener process in  $K$  if

$$W_t = W(t, \cdot) = \sum_{n=1}^{\infty} b_t^n e_n, \quad (5)$$

where  $\{b_t^n\}$  is a sequence of independent, identically distributed (i.i.d.) Brownian motions in one dimension. Then we have

$$EW(t, x) = 0 \quad (6)$$

and

$$\begin{aligned} E(W_t, \theta)(W_s, \varphi) &= E \int_D W(t, x) \theta(x) dx \int_D W(s, y) \varphi(y) dy \\ &= \sum_{n=1}^{\infty} \lambda_n (e_n, \theta)(e_n, \varphi) (t \wedge s) \\ &= (t \wedge s)(Q\theta, \varphi), \end{aligned} \quad (7)$$

where use was made of the fact  $E b_t^m b_s^n = (t \wedge s) \delta_{mn}$ , and  $(t \wedge s) = \min\{t, s\}$ . The Gaussian process  $W_t$  satisfying the properties (6) and (7) is called a  $Q$ -Wiener process in  $K$ .



As an example of stochastic PDE, consider the reaction - diffusion equation with a random drift:

$$\begin{cases} \frac{\partial}{\partial t} u(t, x) = \nu \frac{\partial^2 u}{\partial x^2} + f(u, \frac{\partial u}{\partial x}) - \frac{\partial u}{\partial x} \dot{W}(t, x), & t > 0, x \in D, \\ u(0, x) = \varphi(x) \\ u(t, 0) = u(t, 1) = 0, \end{cases} \quad (8)$$

where  $W_t = W(t, \cdot)$  is a  $Q$ -Wiener process defined as above and  $\dot{W} = \frac{\partial}{\partial t} W$  is formally a  $K$ -valued white noise. Introduce  $H = L^2(D)$  ( $= K$  in this case), with inner product  $(\cdot, \cdot)$  and norm  $\|\cdot\|$ . Let  $V = H_0^1(D) = \{\varphi \in H : \frac{\partial \varphi}{\partial x} \in H \text{ and } \varphi(0) = \varphi(1) = 0\}$  with norm  $\|\cdot\|_1$  and let  $V'$  be the dual space of  $V$ . For  $v \in V$  and  $v^* \in V'$ , the linear functional  $v^*$  evaluated at  $v$  is written as

$$v^*(v) = \langle v^*, v \rangle.$$

Now we define the operator  $A$  and  $B$  as follows:

$$\begin{aligned} A(v) &= \nu \frac{\partial^2 v}{\partial x^2} + f(x, \frac{\partial v}{\partial x}), \\ B(v) &= (-\frac{\partial v}{\partial x}) \cdot \varphi. \end{aligned}$$

If  $f(x, y)$  is bounded and smooth, we have  $A : V \rightarrow V'$  and  $B : V \rightarrow \mathcal{L}^2(K, H)$ , which denotes the space of Hilbert-Schmidt operators from  $K$  into  $H$ . By using the above notations and setting  $u_t = u(t, \cdot)$ , Eq. (8) can be regarded as an Itô equation in  $V'$ :

$$\begin{cases} du_t = A(u_t)dt + B(u_t)dW_t, \\ u_0 = \varphi, \end{cases}$$

or as an integral equation

$$u_t = \varphi + \int_0^t A(u_s)ds + \int_0^t B(u_s)dW_s,$$

where  $\varphi \in H$  and the last integral is a  $H$ -valued Itô integral. The above is a special case of a large class of quasilinear parabolic-Itô equations in a domain  $D \subset \mathbb{R}^d$ , which have been studied by many authors. For references, see Pardoux [6], Krylov and Rozovskii [7], DaPrato and Zabczyk [8].

### III. LYAPUNOV FUNCTIONALS

In general let  $V$  and  $H$  be real separable Hilbert spaces such that  $V \subset H \subset V'$  and the inclusions are dense and compact. By using the same notations as in §II, consider the following Itô equation in  $V'$ :

$$\begin{cases} du_t = A(u_t)dt + B(u_t)dW_t, \\ u_0 = \varphi, \end{cases} \quad (9)$$

where  $A(0) = B(0) = 0$ . Under suitable conditions, the above equation has a unique strong solution  $u_t^\varphi$  satisfying [6]:

$$E \sup_{0 \leq t \leq T} \|u_t^\varphi\|^2 + E \int_0^T \|u_t^\varphi\|_1^p dt < \infty, \text{ for any } T > 0 \text{ and } p \geq 2. \quad (10)$$

As in finite-dimension, the Itô formula plays an important role in stochastic analysis. For a smooth functional  $\Phi$  on  $H$ , let  $D\Phi(v)$  and  $D^2\Phi(v)$  denote the first and the second Fréchet derivatives, respectively. Then the Itô formula associated with Eq. (9) reads

$$\Phi(u_t^\varphi) = \Phi(\varphi) + \int_0^t \mathcal{L}\Phi(u_s^\varphi)ds + \int_0^t (D\Phi(u_s^\varphi), B(u_s^\varphi)dW_s), \quad (11)$$

where

$$\mathcal{L}\Phi(v) = \frac{1}{2} \text{Tr}.[D^2\Phi(v)B(v)QB^*(v)] + \langle A(v), D\Phi(v) \rangle. \quad (12)$$

Let us give two examples:

Ex.1)  $\Phi(v) = \|v\|^2, v \in H.$

Then

$$D\Phi(v) = 2v,$$

$$D^2\Phi(v) = 2I,$$

where  $I$  is the identity operator on  $H$ .

We get

$$\mathcal{L}\Phi(v) = \text{Tr}.[B(v)QB^*(v)] + 2 \langle A(v), v \rangle.$$

Ex.2)  $\Phi(v) = \ell n \|v\|, v \neq 0.$

We have

$$D\Phi(v) = v/\|v\|^2,$$

$$D^2\Phi(v) = (I/\|v\|^2) - (v \otimes v)/\|v\|^4,$$

where  $\otimes$  denotes the tensor product.

Hence

$$\begin{aligned} \mathcal{L}\Phi(v) &= \frac{1}{\|v\|^2} \left\{ \frac{1}{2} \text{Tr}.[B(v)QB^*(v)] + \langle A(v), v \rangle \right\} \\ &\quad - \|Q^{1/2}B^*(v)v\|^2/\|v\|^4. \end{aligned}$$

Let  $\Phi$  be a regular functional on a neighborhood  $U$  of the origin of  $H$ . It is said to be a Lyapunov functional for Eq. (9) if

$$(i) \quad \mathcal{L}(v) \leq 0 \text{ for any } v \in V \cap U \text{ with } v \neq 0, \quad (13)$$

(ii)  $\Phi$  is uniformly positive-definite so that  $\Phi(0) = 0$  and

$$\inf_{\|v\| > r} \Phi(v) = \Phi_r > 0 \quad \forall r > 0. \quad (14)$$

#### IV. STOCHASTIC STABILITY

There are three types of stochastic stability: stability in probability, the  $p$ -th moment stability and the almost sure (a.s.) stability. In what follows, we will consider the first type of stability and the latter two in the asymptotic sense as  $t \rightarrow \infty$ .

We say that the null solution of Eq. (9) is stable in probability if its solution  $u_t^\varphi$  satisfies

$$\lim_{\|\varphi\| \rightarrow 0} P\{\sup_{t > 0} \|u_t^\varphi\| > \varepsilon\} = 0 \quad \text{for any } \varepsilon > 0. \quad (15)$$

The null solution is asymptotically stable in  $p$ -th moment if for any  $\varphi \in U$ , we have

$$\lim_{t \rightarrow \infty} E\|u_t^\varphi\|^p = 0, \quad p > 0, \quad (16)$$

and it is a.s. asymptotically stable if, for any  $\varphi \in U$ ,

$$P\{\lim_{t \rightarrow \infty} \|u_t^\varphi\| = 0\} = 1. \quad (17)$$

The three types of stability as defined by (15)–(17) can be discussed with the aid of a Lyapunov functional. First of all let us assume the existence of such a functional  $\Phi$  in  $U \in H$ . We shall summarize the facts about the stability criteria as three theorems.

**Theorem 1.** The existence of a Lyapunov functional  $\Phi$  in  $U \subset H$  implies that the null solution is stable in probability. #

The proof is a simple consequence of the following Chebyshev inequality:

$$P\{\sup_{0 \leq t \leq T} \|u_t^\varphi\| > r\} \leq \Phi(\varphi)/\Phi_r.$$

In the next theorem, we give some sufficient conditions for the  $p$ -moment stability.

**Theorem 2.** Suppose that there exists a Lyapunov functional in  $H$  such that

$$(i) \quad C_1 \|h\|^p \leq \Phi(h) \leq C_2 \|h\|^p, \text{ for } h \in H, \quad (18)$$

$$(ii) \quad \mathcal{L}\Phi(v) \leq -c_3 \|v\|^p, \text{ for } v \in V, v \neq 0, \quad (19)$$

where  $c_1, c_2, c_3$  are positive constants.

Then the null solution is asymptotically stable in  $p$ -th-moment. #

The proof, based on the Itô formula and the Gronwall inequality, is rather straightforward. For the finite-dimensional case, such a theorem is given by Khasminskii (p.186, [2]). Actually, in this case, the null solution is exponentially stable in  $p$ -th-moment. The last theorem is concerned with a.s. asymptotic stability. For this to be true, intuitively, two things must happen: Starting from any point in  $H$ , the trajectory  $u_t^v$  will be trapped in a ball  $B_r$  of radius  $r$  after some finite time, and, once inside  $B_r$ , the path will eventually find its way to the origin with probability one. The sufficient conditions are given in the following.

**Theorem 3.** Let there exist a Lyapunov functional  $\Phi$  in  $H$  such that

$$(i) \quad \mathcal{L}\Phi(v) \leq -c\Phi(v), \quad (20)$$

for some  $c > 0$  and  $v \in V$ , with  $v \neq 0$ .

$$(ii) \quad \sup_{v \in V} \{ \|Q^{1/2} B^*(v) D\Phi(v)\| / \Phi(v) \} \leq M, \quad (21)$$

where  $M > 0$  is a constant.

Then the null solution is a.s. asymptotically stable. #

The proof of this theorem is more complicated technically, and it can be carried out as in Thm.4.1 of our paper [3] with a minor modification. The conditions (i) and (ii) ensure that the two things mentioned above would indeed happen and the theorem follows.

## V. SOME EXAMPLES.

For the purpose of illustration, let us consider a few examples.

Ex.1). Consider the linear SPDE:

$$\frac{\partial u(t, x)}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} + \dot{W}(t, x)u(t, x), \quad t > 0, \quad 0 < x < 1, \quad (22)$$

subject to the initial-boundary conditions:

$$\begin{cases} u(0, x) = v(x), \\ u(t, 0) = u(t, 1) = 0, \end{cases} \quad (23)$$

where  $\nu$  is a positive constant. In this case, we let  $H = K = L^2(0, 1)$  and let  $H_0^1$  be the Sobolev space as in §II. It was shown by Khasminskii and Mandrekar [5] that the functional

$$\Phi(v) = \int_0^\infty E \|u_t^v\|_1^2 dt \quad (24)$$

is a Lyapunov functional, where

$$\|\varphi\|_1^2 = \|\varphi_x\|^2 + \|\varphi\|^2 = \int_0^1 \{|\varphi'(x)|^2 + |\varphi(x)|^2\} dx.$$

Therefore, by Thm.1, the null solution is stable in probability. In fact the Lyapunov functional (24) satisfies the conditions (18) and (19) of Thm.2 for  $p = 2$ . Thus the null solution is asymptotically stable in second moment or in mean-square. Moreover, if the covariance function  $q(x, y)$  of the  $Q$ -Wiener process  $W(t, x)$  is bounded and continuous, then the conditions (20) and (21) for Thm.3 are met. Hence, in this case the null solution is a.s. asymptotically

stable.

Ex.2). Instead of Eq. (22), we consider the following nonlinear equation:

$$\frac{\partial u(x, t)}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} + \left( \frac{u}{1 + |u|} \right) \dot{W}(t, x), \quad t > 0, \quad 0 < x < 1, \quad (25)$$

with the same initial-boundary conditions (23).

Let

$$\Phi(v) = \|v\|^2. \quad (26)$$

Referring to Eq. (9), we have  $A = \nu \frac{d^2}{dx^2}$  and  $B(v) = \frac{v}{1+|v|}$ . Then it is easy to get

$$\begin{aligned} \mathcal{L}\Phi(v) &= 2 \langle Av, v \rangle + \text{Tr}[B(v)QB^*(v)] \\ &= 2\nu \int_0^1 |v_x(x)|^2 dx + \int_0^1 q(x, x) \frac{v^2(x)}{[1 + |v(x)|]^2} dx \\ &\leq -2\nu \int_0^1 |v_x(x)|^2 dx + \int_0^1 q(x, x) v^2(x) dx. \end{aligned} \quad (27)$$

Therefore, if

$$\int_0^1 q(x, x) v^2(x) dx \leq 2\nu \int_0^1 |v_x|^2 dx,$$

we have

$$\mathcal{L}\Phi(v) \leq 0 \quad \text{for } v \in H_0^1,$$

so that  $\Phi$  given by (26) is a Lyapunov functional and the null solution is stable in probability.

Suppose that  $q$  is bounded and continuous with

$$q_0 = \max_{0 \leq x \leq 1} q(x, x),$$

and

$$\lambda_0 = \inf_{\varphi \in H_0^1} \{ \|\varphi_x\|^2 / \|\varphi\|^2 \} > 0.$$

Then (27) yields

$$\mathcal{L}\Phi(v) \leq (-2\lambda_0 + q_0)\Phi(v).$$

If the following condition holds,

$$q_0 < 2\lambda_0,$$

then it is not difficult to verify that the conditions for both Thm.2 and Thm.3 are fulfilled.

We can conclude that the null solution is asymptotically stable in mean-square as well as with probability one.

Ex.3). Let  $D \subset \mathbb{R}^3$  be a bounded domain with a smooth boundary  $\partial D$ . We consider the reaction-diffusion equation with a random perturbation:

$$\begin{cases} \frac{\partial u(t, x)}{\partial t} = \nu \Delta u + f(u) + \dot{W}(t, x)|\nabla u|, & t > 0, \quad x \in D, \\ u(0, x) = v(x), \\ u|_{\partial D} = 0, \end{cases} \quad (28)$$

where  $f$  is a locally Lipschitz continuous function with at most a polynomial growth and  $f(0) = 0$ . Consider the scalar ordinary differential equation:

$$\frac{dr_t}{dt} = f(r_t),$$

which is assumed to possess a Lyapunov function  $\varphi(r)$  so that  $\varphi''(r) \leq 0$ ,  $\varphi'(r)f(r) \geq 0$  and  $\varphi(r) \rightarrow \infty$  as  $|r| \rightarrow \infty$ .

Let us introduce  $H = K = L^2(D)$ ,  $H_0^1 = \{v \in H : |\nabla v| \in H \text{ and } v|_{\partial D} = 0\}$  and  $V = L^p(D) \cap H_0^1$  for some  $p \geq 2$ . We define

$$\Phi(v) = \int_D \varphi[v(x)]dx. \quad (29)$$



The following calculations are straightforward:

$$\begin{aligned}\mathcal{L}\Phi(v) &= -\nu \int_D \varphi''(v) |\nabla v|^2 dx + \int_D \varphi'(v) f(v) dx \\ &\quad + \frac{1}{2} \int_D \varphi''(v) q(x, x) |\nabla v|^2 dx \\ &\leq - \int_D \varphi''(v) [\nu - \frac{1}{2} q(x, x)] |\nabla v|^2 dx.\end{aligned}$$

If  $q$  is bounded and continuous with

$$q_0 = \sup_{x \in D} q(x, x) \leq 2\nu,$$

then

$$\mathcal{L}\Phi(v) \leq 0 \text{ for } v \in V$$

so that  $\Phi$  defined by (29) is a Lyapunov functional and, by Thm.1, the null solution is stable in probability. But, for this example, the asymptotic stability results are difficult to get with making further assumptions.

#### ACKNOWLEDGEMENT.

This author's work presented here and his research on stochastic partial differential equations in general had been supported by the U.S. Army Research Office over a span of several years around 1980. He wishes to take this opportunity to express his profound gratitude to the ARO for its generous support which has enhanced his research career enormously.

## REFERENCES

1. Gikhman, I.I. and A.V. Skorohod, Stochastic Differential Equations, Springer-Verlag, New York-Berlin-Heidelberg, 1972.
2. Khasminskii, R.Z., Stochastic Stability of Differential Equations, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.
3. Chow, P.L., Stability of nonlinear stochastic evolution equations, J. Math. Anal. and Appl. 89 (1982), 400-419.
4. Ichikawa, A., Semilinear stochastic evolution equations: Boundedness, stability and invariant measure. Stochastics, 12 (1984), 1-39.
5. Khasminskii, R.Z. and V. Mandrekar, On stability of solutions of stochastic evolution equations, preprint (1993).
6. Pardoux, E., Equations aux dérivées partielles stochastiques non linéaires monotones, Thesis, Université Paris XI, 1975.
7. Krylov, N.N. and B.L. Rozovskii, Stochastic evolution equations. J. Soviet Math. 14 (1981), 1233-1277.
8. DaPrato, G. and J. Zabczyk, Stochastic Equations in Infinite Dimensions, Cambridge Univ. Press, Cambridge, England, 1992.

# Analysis and Computation of Approximate Solutions to a Simple Model of Shear Band Formation in One and Two Dimensions

Donald A. French\*  
Department of Mathematical Sciences  
University of Cincinnati  
Cincinnati, Ohio 45221-0025

June 29, 1993

## Abstract

Preliminary results on the analysis and computation of approximate solutions to a simple mathematical model for high strain rate shear deformations of a thermo-plastic material in one and two dimensions are discussed. This problem involves a system of time-dependent partial differential equations. Open questions on the qualitative behavior of the solutions are given and the results of numerical calculations that address these issues are presented. Several time discretization procedures that retain certain fundamental properties of the system of partial differential equations are described. Finally, an optimal order error estimate for a semidiscrete finite element method is furnished.

Presented at the *Eleventh Army Conference on Applied Mathematics and Computing*, Carnegie Mellon University, Pittsburgh, PA, 8-10 June 1993.

**Acknowledgements:** This work was suggested to us by Reza Malek-Madani (United States Naval Academy). We have also benefited from several long discussions with him.

Donald Estep (Georgia Tech) provided useful ideas on the computation of "blowup".

Sonia M.F. Garcia (United States Naval Academy) collaborated with us on the finite element analysis.

## 1 The Shear Band Model:

Consider high strain rate shear deformations of a thermo-plastic material occupying a region  $\Omega \times (-\infty, \infty)$  where  $\Omega = (0, 1)$  for the one-dimensional problem and is a bounded domain in the  $xy$

---

\*Partially funded by the Army Research Office through grant 28535-MA

plane for the two-dimensional problem. We assume the deformations are of antiplane shear type in the two-dimensional case and the applied forces and boundary conditions depend only on  $x \in \Omega$ . Thus, the quantities we seek will also depend only on  $x \in \Omega$  and time,  $t$ . The model we describe is based on the assumption that thermal softening is due to the heat energy generated by the plastic work. We neglect strain hardening and focus on the strain rate effects.

The mathematical problem consists of an equation for balance of momentum

$$(1) \quad \dot{v} - \nabla \cdot \sigma = 0 \text{ on } \Omega, \quad t > 0$$

and balance of heat energy

$$(2) \quad \dot{\theta} - \lambda \Delta \theta = \kappa \sigma \cdot \nabla v \text{ on } \Omega, \quad t > 0$$

where  $\dot{z}$  denotes the time derivative of  $z$ , for simplicity we set density  $\rho = 1$ ,

$$\nabla = \frac{\partial}{\partial x} \text{ or } \nabla = \vec{i} \frac{\partial}{\partial x} + \vec{j} \frac{\partial}{\partial y},$$

$\lambda$  is the thermal conductivity,  $\kappa$  gives the conversion of mechanical energy to heat,  $v = v(x, t)$  is the displacement velocity,  $\sigma = \sigma(x, t)$  is the stress, and  $\theta = \theta(x, t)$  is the temperature. For the constitutive law we take

$$(3) \quad \sigma = \mu(\theta) \nabla v$$

where we consider  $\mu(\theta) = e^{-\alpha\theta}$  for constant  $\alpha > 0$  and  $\mu(\theta) = \theta^\nu$  for constant  $\nu < 0$ . To complete the specification of the initial/boundary value problem we have conditions

$$(4) \quad \sigma \cdot n = \sigma_1 \cdot n \text{ on } \Gamma_s \text{ and } v = v_1 \text{ on } \Gamma_v$$

where  $\partial\Omega = \Gamma_s \cup \Gamma_v$  as well as

$$(5) \quad \theta = 0 \text{ on } \Gamma_D \text{ and } \frac{\partial \theta}{\partial n} = 0 \text{ on } \Gamma_N$$

where  $\partial\Omega = \Gamma_D \cup \Gamma_N$ . Here  $n$  is the outward pointing unit normal to  $\Omega$ ,  $\sigma_1$  and  $v_1$  are given functions. We also need initial conditions

$$(6) \quad v = v_0 \text{ and } \theta = \theta_0 \text{ on } \Omega \text{ at } t = 0.$$

Our goal in this work is to understand the mathematical mechanisms that lead to shear bands. We will study the qualitative aspects of this time dependent model through scientific computations. An important part of this research is to justify the numerical algorithms as much as is possible.

The literature on these problems is extensive. We note [B] for scientific computations on an elastic plastic multi-dimensional model, [DF] for computations with a moving spatial mesh with variable time steps in one dimension, [MM] for a characterization of steady state solutions, [T] for analysis of "blowup", and [W] for scientific computations on a one-dimensional elastic-plastic model.

## 2 Qualitative Behavior of Solutions:

We describe briefly several analytic results for the partial differential equation in this section.

Maddocks and Malek-Madani [MM] give a functional which would be Lyapunov if it was known that the problem (1)-(6) had a unique global solution. We present the short derivation of this property in the one-dimensional case with  $\mu(\theta) = e^{-2\theta}$ ,  $\kappa = \lambda = 1$ , and Dirichlet boundary conditions

$$(7) \quad \theta(0, t) = \theta(1, t) = v(0, t) = v(1, t) = 0, \quad t > 0.$$

Let

$$\sum(\theta, v_x) = \frac{1}{2} e^{-2\theta} v_x^2$$

and note

$$(8) \quad \frac{d}{dt} \sum(\theta, v_x) = -\sigma v_x \dot{\theta} + \sigma \dot{v}_x.$$

Multiplying equation (1) by  $\dot{v}$  and (2) by  $\dot{\theta}$ , adding the resulting equations, integrating with respect to  $x$ , and using integration by parts we have

$$\int_0^1 (\dot{v}^2 + \dot{\theta}^2) dx = -\frac{d}{dt} \int_0^1 \left( \frac{1}{2} \theta_x^2 - \sigma v_x \dot{\theta} \right) dx + \sigma \dot{v}|_0^1.$$

From (8) and the boundary conditions we obtain

$$(9) \quad \frac{d}{dt} I(\theta, v_x) = - \int_0^1 (\dot{v}^2 + \dot{\theta}^2) dx$$

where

$$I(\theta, v_x) = \int_0^1 \left( \frac{1}{2} \theta_x^2 + \sum(\theta, v_x) \right) dx.$$

Tzavaras [T] proves there will be no global solution in the adiabatic ( $\lambda = 0$ ) case. We now give the short "blowup" argument from [T]. We take  $\mu(\theta) = \theta^{-2}$ ,  $\sigma(1, t) = 1$  for  $t > 0$ , and  $\theta(x, 0) = 1$  for  $x \in \Omega$ . Let  $U(\theta) = -\theta^{-1}$  and note  $U' = \mu$ . From (2) with  $\kappa = 1$  we have

$$\dot{\theta} = \sigma v_x.$$

Multiplying by  $\mu(\theta)$  we have

$$\frac{d}{dt} U(\theta) = \mu(\theta) \dot{\theta} = \sigma^2.$$

Integrating from 0 to  $T$ , setting  $x = 1$ , and using the boundary conditions on  $\sigma$  yields

$$U(\theta(1, T)) = U(\theta(1, 0)) + T.$$

From the initial condition and definition of  $U$  we have

$$\theta(1, T) = \frac{1}{1 - T}.$$

Thus the temperature will "blowup" as  $T \rightarrow 1$ . A natural question is: What happens when  $\lambda > 0$ , the non-adiabatic case?

### 3 Numerical Methods:

In this section we describe several numerical methods which have a finite element spatial discretization combined with a finite difference time-step scheme.

Let  $M_h$  be a space of continuous piecewise polynomials of degree  $\leq q - 1$  defined on a "triangulation" of  $\Omega$ . These elements have diameter  $\cong h$ .

For simplicity we assume we are approximating problem (1)-(6) with homogeneous Dirichlet boundary conditions on  $v$  and  $\theta$ . A spatially discrete continuous in time finite element method is: find  $(v_h(\cdot, t), \theta_h(\cdot, t)) \in M_h \times M_h$  such that

$$(\dot{v}_h, \chi) + (\mu(\theta_h) \nabla v_h, \nabla \chi) = 0 \quad \forall \chi \in M_h,$$

$$(\dot{\theta}_h, \varphi) + \lambda(\nabla \theta_h, \nabla \varphi) = \kappa(\mu(\theta_h) |\nabla v_h|^2, \varphi) \quad \forall \varphi \in M_h$$

for  $0 < t < T$  where  $v_h(\cdot, 0) \cong v_0$  and  $\theta_h(\cdot, 0) \cong \theta_0$ . In this method

$$(w, z) = \int_{\Omega} w z \, dA.$$

In [FG] it is shown under the assumption problem (1)-(6) has a unique smooth solution on an interval  $[0, T]$  that

$$\max_{0 \leq t \leq T} \|v(\cdot, t) - v_h(\cdot, t)\| \leq C h^q$$

and

$$\max_{0 \leq t \leq T} \|\theta(\cdot, t) - \theta_h(\cdot, t)\| \leq C h^q$$

where  $C$  is a constant depending on derivatives of  $\theta$  and  $v$  but not on  $h$ . Here  $\|z\|^2 = \int_{\Omega} z^2 \, dA$ . These rates are "optimal" in the sense that they are the best one could expect from the approximation space used.

We now turn to the time discretization. Define  $\text{div}_h : (L^2(\Omega))^d \rightarrow M_h$  where  $d = 1$  on 2 by

$$(\text{div}_h \eta, \chi) = -(\eta, \nabla \chi) \quad \forall \chi \in M_h,$$

$$\Delta_h : M_h \rightarrow M_h \text{ by}$$

$$(\Delta_h \xi, \chi) = -(\nabla \xi, \nabla \chi) \quad \forall \chi \in M_h,$$

and let  $\pi_h$  be the  $L^2$ -projection. With this notation the backward Euler method on the  $n$ th time step is

$$(10) \quad \frac{1}{k}(v_h^{n+1} - v_h^n) - \operatorname{div}_h(\mu(\theta_h^\ell) \nabla v_h^{n+1}) = 0$$

$$(11) \quad \frac{1}{k}(\theta_h^{n+1} - \theta_h^n) - \lambda \Delta_h \theta_h^{n+1} = \pi_h \left( \mu(\theta_h^\ell) |\nabla v_h^m|^2 \right).$$

Here  $k > 0$  is the time step.

The scheme is *fully implicit* if  $\ell = m = n + 1$ . We used this scheme in two dimensions with a fixed point iteration to solve the nonlinear systems and preconditioned conjugate gradients to solve the linear systems. The scheme is called *semi implicit* if  $\ell = m = n$ . We used this scheme in one-dimension with a tridiagonal solver for the linear systems.

Many researchers have investigated time discretization schemes that preserve a discrete energy or Lyapunov functional (see, for instance, [DN], [E], [G], [SS] or [FS]). In one dimension with the list of assumptions used to derive inequality (9) the backward Euler scheme with  $\ell = n + 1$  and  $m = n$  retains a discrete version of this inequality. To see this multiply the first equation by  $v_h^{n+1} - v_h^n$  and integrate with respect to  $x$ . After integration-by-parts noting  $v_h(0, \cdot) = v_h(1, \cdot) = 0$  we have

$$\|v_h^{n+1} - v_h^n\|^2 + k \left( e^{-2\theta_h^{n+1}} v_{h,x}^{n+1}, v_{h,x}^{n+1} - v_{h,x}^n \right) = 0$$

or

$$(12) \quad \|v_h^{n+1} - v_h^n\|^2 + \frac{k}{2} \left[ \left( e^{-2\theta_h^{n+1}} v_{h,x}^{n+1}, v_{h,x}^{n+1} \right) - \left( e^{-2\theta_h^{n+1}} v_{h,x}^n, v_{h,x}^n \right) \right] \\ + \frac{k}{2} \left( e^{-2\theta_h^{n+1}} \left( v_{h,x}^{n+1} - v_{h,x}^n \right), \left( v_{h,x}^{n+1} - v_{h,x}^n \right) \right) = 0.$$

Labeling the term inside the  $[\cdot]$  as  $J$  we have

$$J = \left[ \left( e^{-2\theta_h^{n+1}} v_{h,x}^{n+1}, v_{h,x}^{n+1} \right) - \left( e^{-2\theta_h^n} v_{h,x}^n, v_{h,x}^n \right) \right] - \left( \left( e^{-2\theta_h^{n+1}} - e^{-2\theta_h^n} \right) v_{h,x}^n, v_{h,x}^n \right)$$

Applying Taylor's Theorem we have

$$e^{-2\theta_h^n} = e^{-2\theta_h^{n+1}} + \left( -2e^{-2\theta_h^{n+1}} \right) (\theta_h^n - \theta_h^{n+1}) + \frac{1}{2} \left( 4e^{-2\xi^n} \right) (\theta_h^n - \theta_h^{n+1})^2$$

so

$$J = \left[ \left( e^{-2\theta_h^{n+1}} v_{h,x}^{n+1}, v_{h,x}^{n+1} \right) - \left( e^{-2\theta_h^n} v_{h,x}^n, v_{h,x}^n \right) \right] - \left( 2e^{-2\theta_h^{n+1}} (\theta_h^n - \theta_h^{n+1}) v_{h,x}^n, v_{h,x}^n \right) \\ - \left( -2e^{-2\xi^n} (\theta_h^n - \theta_h^{n+1})^2 v_{h,x}^n, v_{h,x}^n \right)$$

Substituting this back in (12) yields

$$(13) \quad \|v_h^{n+1} - v_h^n\|^2 + \frac{k}{2} \left[ \left( e^{-2\theta_h^{n+1}} (v_{h,x}^{n+1})^2, 1 \right) - \left( e^{-2\theta_h^n} (v_{h,x}^n)^2, 1 \right) \right] \\ + k \left( e^{-2\theta_h^{n+1}} (\theta_h^{n+1} - \theta_h^n) (v_{h,x}^n)^2, 1 \right) \\ + k \left( e^{-2\theta_h^n} (\theta_h^{n+1} - \theta_h^n)^2 (v_{h,x}^n)^2, 1 \right) \\ \frac{k}{2} \left( e^{-2\theta_h^{n+1}} (v_{h,x}^{n+1} - v_{h,x}^n)^2, 1 \right) = 0$$

Multiplying (11) by  $(\theta_h^{n+1} - \theta_h^n)$  we obtain

$$\|\theta_h^{n+1} - \theta_h^n\|^2 + \frac{k}{2} (\|\theta_h^{n+1}\|^2 - \|\theta_h^n\|^2 + \|\theta_h^{n+1} - \theta_h^n\|^2) = k \left( e^{-2\theta_h^{n+1}} (v_{h,x}^n)^2, \theta_h^{n+1} - \theta_h^n \right)$$

Adding this to (13) gives a discrete Lyapunov functional equation for the backward Euler method,

$$\|v_h^{n+1} - v_h^n\|^2 + \|\theta_h^{n+1} - \theta_h^n\|^2 + kE^{n+1} + P = kE^n$$

where

$$E^n = \frac{1}{2} \|\theta_h^n\|^2 + \frac{1}{2} \left( e^{-2\theta_h^n} (v_{h,x}^n)^2, 1 \right)$$

and

$$P = \frac{k}{2} \left[ \left( e^{-2\theta_h^{n+1}} (v_{h,x}^{n+1} - v_{h,x}^n)^2, 1 \right) + \|\theta_h^{n+1} - \theta_h^n\|^2 \right] + k \left( e^{-2\theta_h^n} (\theta_h^{n+1} - \theta_h^n)^2 (v_{h,x}^n)^2, 1 \right).$$

The following scheme also retains a discrete version of (9):

$$\frac{1}{k} (v_h^{n+1} - v_h^n) - \text{div}_h \left( \frac{\sum (\theta_h^{n+1}, v_{h,x}^{n+1}) - \sum (\theta_h^{n+1}, v_{h,x}^n)}{v_{h,x}^{n+1} - v_{h,x}^n} \right) = 0 \\ \frac{1}{k} (\theta_h^{n+1} - \theta_h^n) - \frac{1}{2} \Delta_h (\theta_h^{n+1} + \theta_h^n) = -\pi_h \left( \frac{\sum (\theta_h^{n+1}, v_{h,x}^n) - \sum (\theta_h^n, v_{h,x}^n)}{\theta_h^{n+1} - \theta_h^n} \right).$$

Multiplying the first equation by  $k^{-1}(v_h^{n+1} - v_h^n)$  and the second by  $k^{-1}(\theta_h^{n+1} - \theta_h^n)$  gives the following after adding the two equations and integrating with respect to  $x$ :

$$\frac{1}{2} (\|\theta_h^{n+1}\|^2 - \|\theta_h^n\|^2) + \int_0^1 \left( \sum (\theta_h^{n+1}, v_{h,x}^{n+1}) - \sum (\theta_h^n, v_{h,x}^n) \right) dx \\ = -k \left( \|k^{-1} (v_h^{n+1} - v_h^n)\|^2 + \|k^{-1} (\theta_h^{n+1} - \theta_h^n)\|^2 \right)$$

Finally, we note that higher order accurate schemes can be created using the finite element in time techniques described in [FS].



## 4 Numerical Experiments:

In this section we show the results of some of our preliminary computations. More experiments are planned and we intend to use more sophisticated adaptive methods.

Figure 1 shows the results of a computation in the adiabatic case ( $\lambda = 0$ ). As discussed earlier [T] has shown that the temperature,  $\theta$ , will "blowup" at the  $x = 1$  boundary as  $t \rightarrow 1$ . We take

$$\mu(\theta) = \theta^{-2}, \quad \sigma_1 = 1, \quad v_0(x) = \frac{1}{2}x^2, \quad \text{and} \quad \theta_0(x) = 1$$

A refined mesh and variable time step strategy were used. The top graph has a snapshot of  $\theta_h$  at  $T = 0.9$ . The bottom graph has the evolution of  $\theta_h$  at  $x = 1$  plotted against its known values  $(1 - t)^{-1}$ .

Figure 2 has the results from nearly the same calculation except we took  $\lambda = 0.1$ , added Neumann boundary conditions for  $\theta$ , and stopped at  $T = 1.25$ . In this case it can be shown that

$$\theta(1, t) = \left( 1 - \int_0^t \mu(\theta) \theta_{xx} d\tau \right)^{-1}.$$

So, to estimate the behavior of  $\theta_h(1, t)$  we used a centered difference approximation of " $\theta_{h,xx}$ ". These are the values on the dashed line in the bottom graph. Certainly from this numerical evidence it appears that the solution will "blowup".

Finally, in figure 3 we explore "blowup" in two dimensions. We have the adiabatic case ( $\lambda = 0$ ),  $\mu(\theta) = \theta^{-2}$ ,  $v_0(x, y) = \frac{1}{2}(1 - x^2)y^2$ , and  $\theta_0(x, y) = 1$ . The domain is  $\Omega = (-1, 1) \times (0, 1)$ . We took  $v = 0$  and  $\theta = 1$  on all boundaries except the segment of the line  $y \equiv 1$ . On that boundary we had  $\frac{\partial \theta}{\partial n} = 0$  and  $\sigma \cdot n = (1 - x^2)$ . We used a  $40 \times 40$  refined mesh and a variable time step with 400 steps. Again we see a possible blowup point starting to form.

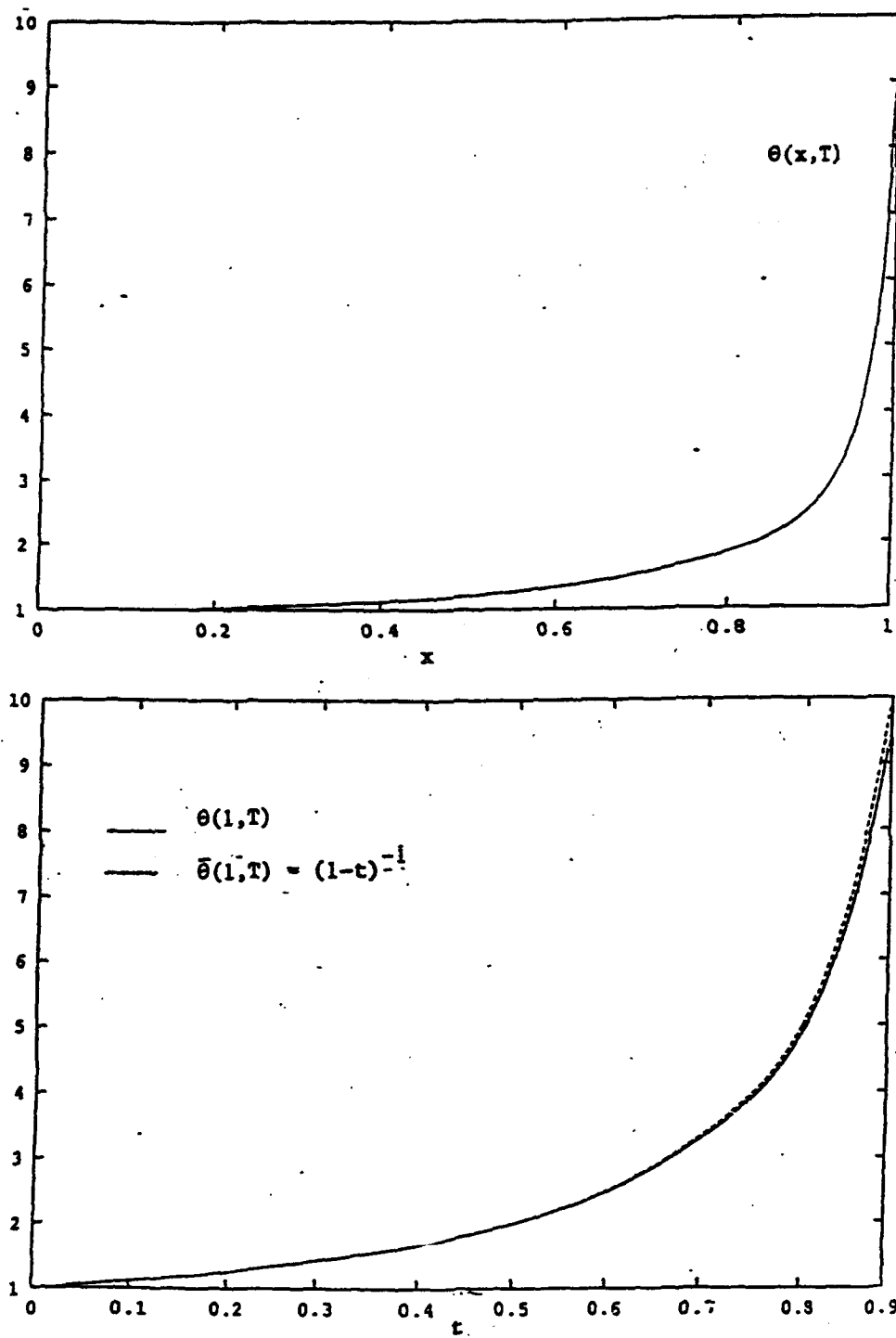


Figure 1

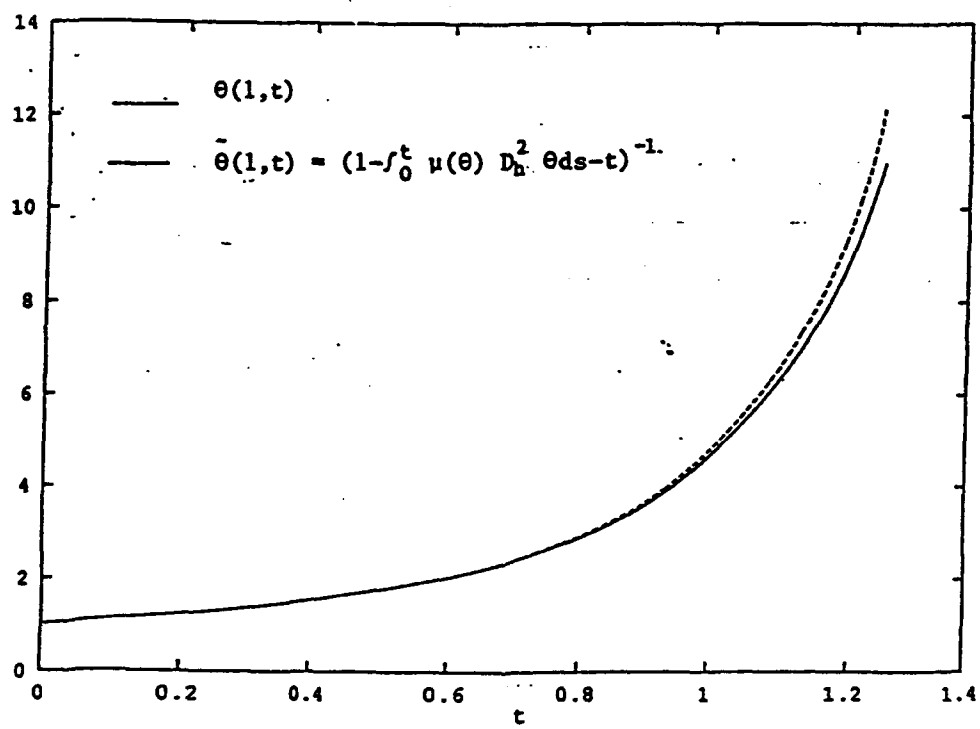
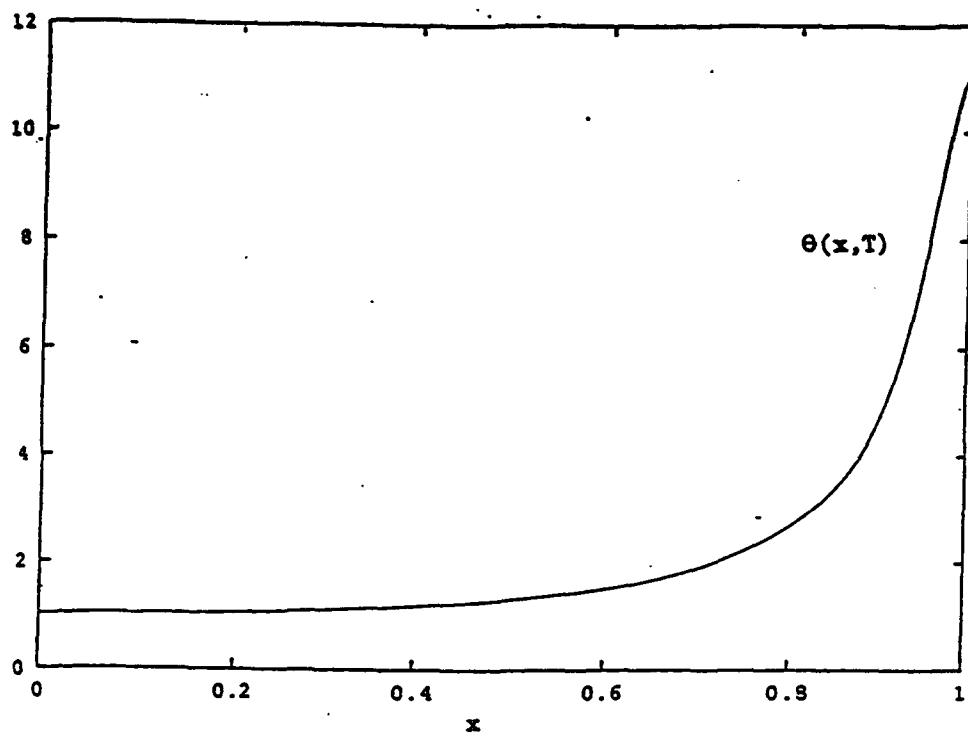


Figure 2

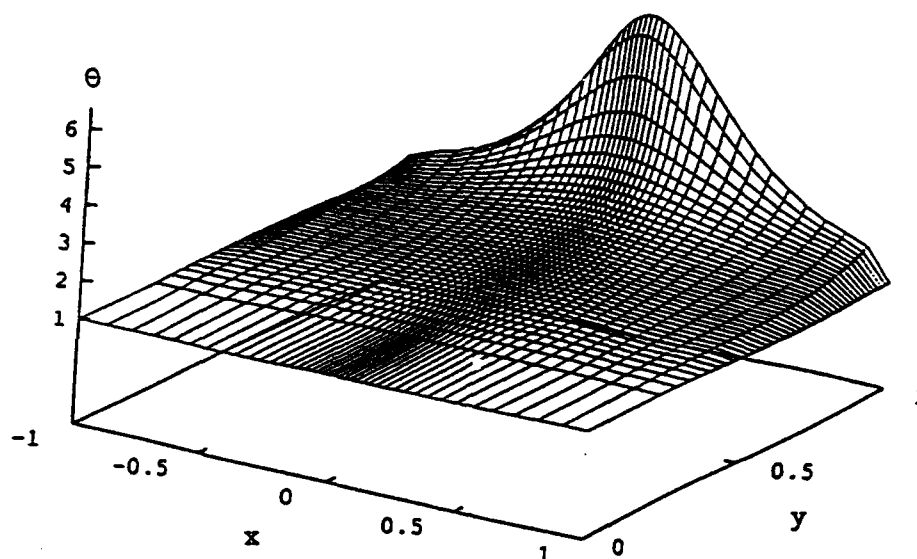


Figure 3

## 5 Specialized Finite Difference Method:

We finish this report by describing a finite difference scheme for the adiabatic problem that has the same "blowup" as described earlier from [T]. The scheme will give approximations to

$$(14) \quad \dot{v} = \sigma_x \text{ on } (0,1), \quad t > 0,$$

$$(15) \quad \dot{\theta} = \mu(\theta)v_x^2 \text{ on } (0,1), \quad t > 0,$$

with boundary conditions

$$\sigma(0,t) = 0, \quad \sigma(1,t) = 1, \quad t > 0$$

where  $\sigma = \mu(\theta)v_x$  and initial conditions are provided for  $\theta$  and  $v$ .

Let

$$\theta_\ell^n \cong \theta(\ell h, nk)$$

$$v_{\ell+1/2}^n \cong v((\ell + 1/2)h, nk),$$

$$\bar{\mu}(\varphi, \psi) = \frac{U(\varphi) - U(\psi)}{\varphi - \psi}, \quad U' = \mu,$$

and

$$\sigma_\ell^{n+1/2} = \bar{\mu}(\theta_\ell^{n+1}, \theta_\ell^n) \frac{v_{\ell+1/2}^{n+1/2} - v_{\ell-1/2}^{n+1/2}}{h}$$

where

$$v_{\ell+1/2}^{n+1/2} = \frac{1}{2} (v_{\ell+1/2}^{n+1} + v_{\ell+1/2}^n).$$

The scheme for (14-15) is then

$$\frac{1}{k} (v_{\ell+1/2}^{n+1} - v_{\ell+1/2}^n) = \frac{1}{h} (\sigma_{\ell+1}^{n+1/2} - \sigma_\ell^{n+1/2}), \quad \ell = 0, 1, \dots, L-1,$$

$$\frac{1}{k} (\theta_\ell^{n+1} - \theta_\ell^n) = \bar{\mu}(\theta_\ell^{n+1}, \theta_\ell^n) (h^{-1} (v_{\ell+1/2}^{n+1/2} - v_{\ell-1/2}^{n+1/2}))^2, \quad \ell = 0, 1, \dots, L$$

and

$$\sigma_0^{n+1/2} = 0, \quad \sigma_h^{n+1/2} = 1.$$

We now show that the approximation,  $\theta_h^n$ , will "blowup". Take  $\mu(\varphi) = \varphi^{-2}$  so  $U(\varphi) = -\varphi^{-1}$  and set  $\theta_\ell^0 = 1$  for  $\ell = 0, 1, \dots, L$ . Multiplying the second equation by  $\bar{\mu}(\theta_\ell^{n+1}, \theta_\ell^n)$  we have

$$\bar{\mu}(\theta_\ell^{n+1}, \theta_\ell^n) (\theta_\ell^{n+1} - \theta_\ell^n) = k (\sigma_\ell^{n+1/2})^2.$$

At  $\ell = L$  we have  $\sigma_L^{n+1/2} = 1$  so this becomes

$$U(\theta_h^{n+1}) - U(\theta_h^n) = k.$$

Iterating this formula we have

$$(\theta_h^N)^{-1} = (\theta_h^0)^{-1} - Nk$$

or

$$\theta_h^N = \frac{1}{1 - Nk}$$

which proves the "blowup".

## References

- [B] R.C. Batra, Analysis of shear bands in simple shearing deformations of nonpolar and dipolar viscoplastic materials, *Appl. Mech. Rev.*, **45**(1992), 123-131.
- [DF] D.A. Drew and J.E. Flaherty, Adaptive finite element methods and the numerical solution of shear band problems, in *Phase Transformations and Material Instabilities in Solids*, Academic Press (1984), 37-60.

- [DN] Q. Du and R.A. Nicolaides, Numerical analysis of a continuum model of phase transition, *SIAM J. Num. Anal.*, **28**(1984), 1310-1322.
- [E] C.M. Elliott, The Cahn-Hilliard model for the kinetics of phase separation, in *Mathematical Models for Phase Change Problems*, J.F. Rodrigues, Birkhauser Verlag (1989).
- [FG] D.A. French and S.M.F. Garcia, Finite element approximation of an evolution problem modeling shear band formation (Submitted to *Comp. Meth. Appl. Mech. Eng.*).
- [FS] D.A. French and J.W. Schaeffer, Continuous finite element methods which preserve energy properties for nonlinear problems, *Appl. Math. Comp.*, **39**(1990), 271-295.
- [G] R.T. Glassy, Convergence of an energy preserving scheme for the Zakharov equations in one space dimension, *Math. Comp.*, **58**(1992), 83-102.
- [MM] J.H. Maddocks and R. Malek-Madani, Steady-state shear-bands in thermo-plasticity. I: Vanishing yield stress, *Int. J. Solids Structures*, **29**(1992), 2039-2061.
- [SS] J.M. Sanz-Serna, Methods for the numerical solution of the nonlinear Schrödinger equation, *Math. Comp.*, **43**(1984), 21-27.
- [T] A.E. Tzavaras, Effect of thermal softening in shearing of strain-rate dependent materials, *Arch. Rat. Mech. Anal.*, **99**(1987), 349-374.
- [W] J.W. Walter, Jr., Numerical experiments on adiabatic shear band formation in one dimension, *BRL Technical Report BRL-TR-3381* (1991), 1-67.

# A System for Materials of Korteweg Type in Two Space Variables

Harumi Hattori\* and Dening Li†  
Department of Mathematics  
West Virginia University

## 1 Viscous Isothermal Motion of Korteweg Type Materials.

We consider a simplified isothermal motion of the Korteweg type materials where the viscous effect is included. The general model for the Korteweg type materials was proposed by Dunn and Serrin [1] with the interstitial working term and they modified the system of compressible fluids based on the Korteweg theory of capillarity. We will restrict ourselves to the special isothermal case.

With the Helmholtz free energy chosen as

$$\psi = F(\rho) + \frac{\nu}{2\rho}(\rho_x^2 + \rho_y^2), \quad (1.1)$$

where  $F$  is a smooth function of  $\rho$  and  $\nu$  is a positive constant, we obtain the following system for

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2)_x + (\rho uv)_y = -p_x + \nu \rho \Delta \rho_x + \mu \Delta u + \frac{\mu}{3}(u_{xx} + v_{xy}), \\ (\rho v)_t + (\rho uv)_x + (\rho v^2)_y = -p_y + \nu \rho \Delta \rho_y + \mu \Delta v + \frac{\mu}{3}(u_{xy} + v_{yy}). \end{cases} \quad (1.2)$$

with the initial data

$$(\rho, u, v)(x, y, 0) = (\rho_0, u_0, v_0)(x, y). \quad (1.3)$$

---

\*The work was supported in part by Army Grant DAAL 03-89-G-0088

†The work was supported in part by ONR under Grant N00014-91-J-1291.

Where  $p' > 0$ ,  $p'' > 0$ :

$$H'(\rho) = h(\rho) = \int_{\bar{\rho}_0}^{\rho} \frac{p'(\rho)}{\rho} d\rho, \quad (1.4)$$

$$H(\bar{\rho}_0) = 0, \quad \bar{\rho}_0 = \text{const} > 0. \quad (1.5)$$

$$H(\rho) \geq \delta(\rho - \bar{\rho}_0)^2. \quad (1.6)$$

## 2 Local Solution

First we discuss the local existence of solutions for the initial value problem of (1.2)(1.3).

Denote by  $\|\cdot\| \equiv \|\cdot\|_0$  the  $L^2$  norm and by  $\|\cdot\|_k$  the  $k$ -th order Sobolev norm. Set

$$\|w\|_{0,T}^2 \equiv \sup_{0 \leq t \leq T} (\|w(t)\|^2 + \|\nabla \rho(t)\|^2) + \int_0^T (\|\nabla u(t)\|^2 + \|\nabla v(t)\|^2) dt \quad (2.1)$$

and

$$\|w\|_k^2 = \sum_{|j| \leq k} \|\partial_{x,y}^j w\|^2, \quad (2.2)$$

where  $w \equiv (\rho, u, v)$ . Then the main result of the local existence is the following theorem

**Theorem 1** Assume the initial data (1.3) satisfy

$$(\rho_0 - \bar{\rho}_0, u_0, v_0) \in H^k(R^2), \quad \bar{\rho}_0 \geq \delta > 0, \quad (2.3)$$

where  $k \geq 4$  and  $\bar{\rho}_0 > 0$  is a positive constant.

Then,  $\exists T > 0$  such that Cauchy problem (1.2)(1.3) has a unique solution  $w \equiv (\rho - \bar{\rho}_0, u, v)$  in  $[0, T]$  such that

1.  $\rho - \bar{\rho}_0 \in L^\infty([0, T]; H^{k+1}(R^2)),$

2.  $(u, v) \in L^\infty([0, T]; H^k(R^2)),$

- 3.

$$\|w\|_k^2 \leq C_k (\|w_0\|_k^2 + \|\rho_0\|_{k+1}^2).$$

Where

$$\begin{aligned} \|w\|_0^2 &\equiv \sup_{0 \leq t \leq T} (\|w(t)\|^2 + \|\nabla \rho(t)\|^2) \\ &\quad + \int_0^T (\|\nabla u(t)\|^2 + \|\nabla v(t)\|^2) dt, \end{aligned} \quad (2.4)$$

$$\|w\|_k^2 = \sum_{|j| \leq k} \|\partial_{x,y}^j w\|_0^2.$$



The theorem is proved by establishing the energy estimate for the linearized problem and linear iteration.

Since the linearized problem of (1.2) is not of any classical type, the existence of solutions is not known even for the linearized problem. We prove the existence of solutions for the linearized problem by establishing an energy estimate for the dual problem and then using the dual argument.

The linearization of the problem (1.2)(1.3) is the following

$$\begin{cases} (\partial_t + u\partial_x + v\partial_y)\dot{\rho} + \rho(\dot{u}_x + \dot{v}_y) = \dot{f}_1, \\ \rho(\partial_t + u\partial_x + v\partial_y)\dot{u} + p'(\rho)\dot{\rho}_x \\ \quad - \nu\rho\Delta\dot{\rho}_x - \mu\Delta\dot{u} - \frac{\mu}{3}(\dot{u}_{xx} + \dot{v}_{xy}) = \dot{f}_2, \\ \rho(\partial_t + u\partial_x + v\partial_y)\dot{v} + p'(\rho)\dot{\rho}_y \\ \quad - \nu\rho\Delta\dot{\rho}_y - \mu\Delta\dot{v} - \frac{\mu}{3}(\dot{u}_{xy} + \dot{v}_{yy}) = \dot{f}_3. \end{cases} \quad (2.5)$$

$$(\dot{\rho}, \dot{u}, \dot{v})(x, y, 0) = (\dot{\rho}_0, \dot{u}_0, \dot{v}_0)(x, y). \quad (2.6)$$

For the linear problem (2.5)(2.6), we have the estimate

$$\begin{aligned} & \partial_t(\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2) + \|\dot{u}\|_{k+1}^2 + \|\dot{v}\|_{k+1}^2 \\ & \leq C_k(\|\dot{w}\|_k^2 + \|\dot{\rho}\|_{k+1}^2 + \|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2) \\ & \|\dot{w}\|_k^2 \leq C_k(T)(\|\dot{w}_0\|_k^2 + \|\dot{\rho}_0\|_{k+1}^2) \\ & \quad + C_k(T) \int_0^T (\|\dot{f}\|_k^2 + \|\dot{f}_1\|_{k+1}^2) dt. \end{aligned} \quad (2.7)$$

The estimate (2.7) is derived by integrating by parts the inner products of the last two components of (2.5) with  $(\dot{u}, \dot{v})$ , and using repeatedly the conservation of mass equation to treat the term  $\rho(\dot{u}_x + \dot{v}_y)$  coming from the non-symmetry of the system.

For the existence of solutions for linear problem (2.5)(2.6), we need to consider the dual problem

$$L^*\dot{\phi} = -\partial_t\dot{\phi} - B_1^*\partial_x\dot{\phi} - B_2^*\partial_y\dot{\phi} + (T_1^* + T_2^*)\dot{\phi} = \dot{g}, \quad (2.8)$$

$$\dot{\phi}(x, y, T) = 0. \quad (2.9)$$

System (2.8) is non-symmetric with the principal part:

$$\begin{cases} \partial_t\dot{\phi}_1 - \Delta(\partial_x\dot{\phi}_2 + \partial_y\dot{\phi}_3) = \dot{g}_1, \\ \partial_t\dot{\phi}_2 + \partial_x\dot{\phi}_1 + \Delta\dot{\phi}_2 + \partial_{xy}\dot{\phi}_3 = \dot{g}_2, \\ \partial_t\dot{\phi}_3 + \partial_y\dot{\phi}_2 + \Delta\dot{\phi}_3 + \partial_{xy}\dot{\phi}_2 = \dot{g}_3, \end{cases} \quad (2.10)$$

For the system (2.8)(2.9), we have the dual estimate:

$$\|\dot{\phi}(t)\|^2 + \|\dot{\phi}_{2,3}(t)\|_2^2 \leq C \left( \|\dot{g}(t)\|^2 + \int_0^t \|\dot{g}(\tau)\|^2 d\tau \right) \quad (2.11)$$

and the negative order estimate:

$$\|\Lambda^s \dot{\phi}(t)\|^2 + \|\Lambda^{s+2} \dot{\phi}_{2,3}(t)\|^2 \leq C \left( \|\Lambda^s \dot{g}(t)\|^2 + \int_0^T \|\Lambda^s \dot{g}(\tau)\|^2 d\tau \right). \quad (2.12)$$

where  $s \in R$  and  $\Lambda$  is the operator with symbol

$$\lambda(\xi, \eta) = \sqrt{1 + |\xi|^2 + |\eta|^2}. \quad (2.13)$$

The estimate (2.12) is obtained by integrating by parts the inner product of 2nd and 3rd equations with  $(\dot{\phi}_2, \dot{\phi}_3)$  and  $(\Delta \dot{\phi}_2, \Delta \dot{\phi}_3)$ . The negative norm estimate (2.13) is derived similarly by applying the operator  $\Lambda^s$  to the system. Once we have the dual estimate, the existence of solutions for (2.5)(2.6) can be established by the standard dual argument.

From the existence of solutions for the linearized problem (2.5)(2.6) and their estimate (2.7), the solution for the nonlinear problem (1.2)(1.3) is obtained by linear iteration. First, we construct approximate solution  $\tilde{w}(x, y, t)$  such that:

$$\partial_t \tilde{w} - \Delta \tilde{w} = 0, \quad \tilde{w}(x, y, 0) = w_0(x, y) \quad (2.14)$$

so that the Cauchy problem (1.2)(1.3) is transformed into a problem with homogeneous initial data:

$$\begin{cases} \dot{L}(\tilde{w})\tilde{w} = -\mathcal{L}(\tilde{w})\tilde{w} \equiv \tilde{f}, \\ \tilde{w}(x, y, 0) = 0. \end{cases} \quad (2.15)$$

Take  $\tilde{w}_0 = 0$  to begin the iteration scheme. The successive  $\tilde{w}_{j+1}$  ( $j = 0, 1, \dots$ ) is decided by solving the following problems

$$\begin{cases} \dot{L}(\tilde{w}_j)\tilde{w}_{j+1} = \tilde{f}, \\ \tilde{w}_{j+1}(x, y, 0) = 0. \end{cases} \quad (2.16)$$

The existence of the iteration sequence in a common interval is derived from the fact that the norm  $\|\tilde{w}_j\|_k^2$  is uniformly bounded. And the convergence of the iteration sequence is guaranteed by choosing  $T \ll 1$  such that

$$\|\tilde{w}_j\|_k^2 \ll 1, \quad (2.17)$$

$$\|\tilde{w}_j - \tilde{w}_{j-1}\|_{k-2} \leq \frac{1}{2} \|\tilde{w}_{j-1} - \tilde{w}_{j-2}\|_{k-2}. \quad (2.18)$$

Consequently, the limit of the iteration sequence is the desired solution. The uniqueness of the solution is derived readily from the energy estimate.

### 3 Classical Global Solution

The existence of global solution for (1.2)(1.3) can be obtained from the local existence theorem in section 2 by using the Matsumura-Nishida technique in [3]. We have the following

**Theorem 2** Assume that

$$(\rho_0 - \bar{\rho}_0, \nabla \rho_0, u_0, v_0) \in H^4(R^2), \quad \bar{\rho}_0 \geq \delta > 0. \quad (3.1)$$

$$\|w_0\|_4^2 + \|\rho_0 - \bar{\rho}_0\|_5^2 \leq \epsilon_0. \quad (3.2)$$

Then, for  $\epsilon_0 \ll 1$ , there exists a unique solution  $(\rho, u, v)$  in  $[0, \infty)$  such that

$$\rho - \bar{\rho}_0 \in L^\infty([0, \infty); H^5(R^2)), \quad (u, v) \in L^\infty([0, \infty); H^4(R^2)), \quad (3.3)$$

and satisfying

$$\|w\|_4^2 \leq C (\|w_0\|_4^2 + \|\rho_0 - \bar{\rho}_0\|_5^2) \quad (3.4)$$

with  $w \equiv (\rho - \bar{\rho}_0, u, v)$ .

The theorem is proved by applying the following lemma similar to the one in [3].

**Lemma 1** Let  $w$  be a solution such that

$$\sup_{0 \leq t \leq T} (\|w(t)\|_4 + \|\rho(t) - \bar{\rho}_0\|_5) \leq \epsilon. \quad (3.5)$$

If  $\epsilon$  is small, then

$$\sup_{0 \leq t \leq T} (\|w(t)\|_4^2 + \|\rho(t) - \bar{\rho}_0\|_5^2) \leq C_\epsilon (\|w_0\|_4^2 + \|\rho_0 - \bar{\rho}_0\|_5^2). \quad (3.6)$$

Here  $C_\epsilon$  is independent of  $T$ .

The inequality (3.6) can be derived by looking carefully the derivation of the inequality (2.7). We will need the assumption that  $p' > 0$ ,  $p'' > 0$  and (1.4)(1.6), as well as the fact that the "error" terms in the linearization are of quadratic order. Once (3.6) is established, the existence of the global solution follows readily from the standard continuation argument.

### References

1. J.E. Dunn and J. Serrin, *On the thermodynamics of interstitial working*, Arch. Rat. Mech. Anal. 88 (1985), 95-133.
2. D.J. Korteweg, *Sur la forme que prennent les équations des mouvement des fluides si l'on tient compte des forces capillaires par des variations de densité*, Arch. Neerl. Sci. Exactes. Nat. Ser. II 6 (1901), 1- 24.
3. Matsumura and Nishida, *The initial vallue problem for the equations fo motion of viscous and heat-conductive gases*, J. Math. Kyoto Univ. 20(1980), 67-104.
4. J. Serrin, *The form of interfacial surfaces in Korteweg's theory of phase equilibria*, Quart. J. Appl. Math. 41 (1983), 357-364.
5. J. Serrin, *Phase transition and interfacial layers for van der Waals fluids*, in "Proceedings of SAFA IV Conference, Recent Methods in Nonlinear Analysis and Applications, Naples, 1980" (A. Camfora, S. Rionero, C. Sbordone, C. Trombetti, Eds.)

# DYNAMICS OF A POLYMER MOLECULE NEAR A SINGLE STREAMWISE VORTEX

Joseph D. Myers  
Department of Mathematical Sciences  
United States Military Academy  
West Point, New York 10996

September 20, 1993

## Abstract

The addition of a few parts per million (by weight) of certain long-chained polymer molecules dramatically reduces the drag of a turbulent flow, often by a factor of about 3 or 4. This phenomenon is known as Toms effect. It has been tested and exploited in applications ranging from oil pipelines to noise reduction of submarines, yet little is understood about the physics of this phenomenon. It is generally surmised that turbulent flows "somehow" stretch the polymers, thereby increasing the viscosity locally, thickening the viscous sublayer, and thus reducing the velocity gradient at the boundary. What is lacking is the mechanism by which large scale turbulent structures can stretch individual polymer molecules, a problem spanning several orders of magnitude.

In this paper, we adopt a hybrid ellipsoid-dumbbell model to model the dynamics of a polymer molecule. We embed these model molecules within a numerical simulation of a single streamwise vortex in a shear flow. The effect of the vortex is to rotate the flow gradient within the vortex core, thereby establishing inflectional velocity profiles; we also find that it establishes regions of positive streamwise strain rate on the downflow side of the vortex. These regions are observed to be of sufficient strength to stretch the polymer molecule to experimentally observed elongations. Regions of negative strain rate on the upflow side of the vortex are observed to allow the molecules to relax, and admit the possibility of entanglement. We estimate the local increase in viscosity due to polymer deformation, and use this to infer the local decrease in strength of the associated vortex, the reduced inflectional velocity profiles within the vortex core, and the increased stability of the flow locally.

## Contents

1	Introduction	2
2	Modeling the Dynamics of a Single Polymer Molecule	2
3	Benchmark: Comparison to Previous Results	8
4	Dynamics of the Embedded Polymer	12
5	Implications for Flow Stability	20

## 1 Introduction

It is widely suspected that turbulence suppression and turbulent drag reduction by dilute polymer additives occur because the polymers dissipate energy via stretching. However, previous investigations of the dynamics of polymers in uniform shear flow have shown that the embedded polymers undergo too little deformation to dissipate any significant amount of energy. In this paper, we identify a physical mechanism in a single streamwise vortex flow that is capable of substantially deforming polymers, thereby dissipating enough energy to help stabilize the flow and reduce drag. We identify this mechanism by modeling the dynamics of a polymer molecule which has been embedded in a streamwise dependent vortical flow. We investigate how the flow establishes regions with positive and negative streamwise strain rates of sufficient strength to stretch the polymer to experimentally observed deformations, and then to permit relaxation, possibly with subsequent entanglement. We also discuss the resulting predictions of the single vortex model: the increase in viscosity obtained from energy dissipation about the stretched polymer decreases the vortex strength  $R_\nu$  locally near those vortices which have activated polymers, thus reducing the inflectional profiles at the vortex cores, and so preventing those vortices from undergoing transition to turbulence.

## 2 Modeling the Dynamics of a Single Polymer Molecule

We first present a model for the dynamics of a single polymer molecule in a 3-D flow field. In a previous work, Keyes and Abernathy (1987) successfully employed a hybrid ellipsoid-dumbbell polymer model in uniform shear flow to explain observed "early turbulence" in dilute polymer shear flows as being polymer-induced velocity fluctuations, with decreased frequency in regions of increased shear. In this model, the polymer is viewed as simultaneously having a dual nature. When considering deformational effects, the polymer is viewed as a bead-and-spring dumbbell; specifically, as two spheres of equal radius connected by a nonbendable spring that does not interact with the surrounding flow. When considering rotational effects, the polymer is viewed as a rigid prolate ellipsoid whose interior consists of the polymer and the accompanying entrained fluid (Figure 1). This dual-natured model derives from the following observations. Self-avoiding random walk studies show that the 3-D conformation of a random coil polymer is contained within an envelope with a shape similar to that of a rounded bar of soap. The characteristic lengths of this envelope occur roughly in the ratio 3:1.5:1 (Rubin and Mazur, 1975). In a uniform shear flow, polymers exhibit a single response time. In low strain rate flows, the polymers appear to undergo motions as though they were rigid objects. Frequency measurements indicate that this motion is similar to the convection and flipping of solid ellipsoids. In higher strain rate flows, the flipping frequency changes in such a way as to indicate that the polymers have deformed. The simplest single time constant model that will capture this deformational behavior is that of two massless dumbbell ends connected by a spring. Spring deformation is determined by the balance between the spring force and the surface drag force on the two dumbbell ends. The rotational modeling remains that of a rigid ellipsoid. No attempt is made to capture the internal flow within the polymer envelope. Rather, the model is designed to model the motions of the polymer molecule both at low strain rates and at strain rates which are high enough to cause some deformation. We employ the Keyes-Abernathy polymer model here, generalizing it to 3-D flows with streamwise velocity gradients. In the following, all quantities have been nondimensionalized with respect to the usual characteristic values, unless stated otherwise.

Consider a polymer molecule embedded in a local flow field  $\vec{u} \equiv \langle u, v, w \rangle$ . We begin by defining a local Cartesian coordinate system within the frame of the polymer. Let the origin be the instantaneous location of the center of the polymer at the beginning of the computational time step. Choose one axis in the flow direction,  $\vec{u}$ . Choose another axis  $\vec{s}$  to be the component of the shear vector

perpendicular to  $\vec{u}$ :  $\vec{s} \equiv \nabla|\vec{u}| - \hat{u} \cdot \nabla|\vec{u}| \hat{u}$ . The third axis is then  $\vec{f} \equiv \vec{u} \times \vec{s}$ . Normalize each to obtain the flow axis  $\hat{u}$ , the shear axis  $\hat{s}$ , and the flipping axis  $\hat{f}$ . By construction, these three vectors are orthogonal. In this coordinate system, the dimensionless local shear  $\kappa \equiv \nabla|\vec{u}| \cdot \hat{s}$  will cause the polymer to flip about the  $\hat{f}$  axis. Local flow velocities work against the stresses internal to the polymer to translate the polymer ends, usually at different rates. Deformations of the polymer due to shear stress and strain occur in the direction of the ellipsoid's semi-major axis; we neglect bending of the polymer molecule. We assume the two effects of rotation and deformation to be separable; we calculate them independently at each step, and then superpose them to find the resultant motion of the molecule.

We first consider the flipping motion of the polymer. Toward this purpose, we envision the polymer as a rigid prolate ellipsoid whose semi-major and semi-minor axes have lengths  $A$  and  $B$  respectively, where  $A$  and  $B$  are both measured in units of  $\mathcal{H}$ . Let  $\vec{L}_i$  be the dimensional position vector from the polymer center to a point on the semi-major axis which is a distance  $B\mathcal{H}$  short of the  $i$ 'th tip of the polymer, where  $i = 1$  or  $2$  for the two ends of the polymer. (This point serves as the bead center and spring end in the bead-and-spring dumbbell formulation below.) We nondimensionalize the polymer's physical dimensions with respect to the dimensional length of the semi-minor axis,  $B\mathcal{H}$ . The dimensionless position vector is  $\vec{\xi}_i = \frac{\vec{L}_i - \vec{L}_{3-i}}{2B\mathcal{H}}$ . Let  $\vec{\xi}$  denote the  $\vec{\xi}_i$  with positive  $\hat{s}$  component. Note that  $\vec{\xi}$  is parallel to the polymer's semi-major axis. A convenient dimensionless measure of polymer shape is  $\xi \equiv |\vec{\xi}| = \frac{A}{B} - 1$ . Denote the projection of  $\vec{\xi}$  into the  $\hat{u}$ - $\hat{s}$  plane by  $\vec{\xi}_p$ , with magnitude  $\xi_p = |\vec{\xi}_p|$ . Let  $\phi$  denote the angle in the  $\hat{u}$ - $\hat{s}$  plane from the shear axis  $\hat{s}$  to  $\vec{\xi}_p$ . The local shear  $\kappa$  causes the polymer to flip about the  $\hat{f}$  axis in the local  $\hat{u}, \hat{s}, \hat{f}$  coordinate system at the polymer center. Jeffery (1922) has calculated the angular velocity of this flipping motion by setting the net torque on the ellipsoid equal to zero, thereby obtaining:

$$\frac{d\phi}{dt} = \kappa \frac{(\xi_p + 1)^2 \cos^2 \phi + \sin^2 \phi}{(\xi_p + 1)^2 + 1}. \quad (1)$$

Note from this equation that a sphere ( $\xi_p = 0$ ) rotates at constant angular velocity as it is convected along. As the aspect ratio  $\xi_p$  increases, the angular velocity is slowest when the semi-major axis is aligned with the flow ( $\phi = \frac{\pi}{2}$ ) and is fastest when the ellipsoid is broadside to the flow direction ( $\phi = 0$ ), that is, when the semi-minor axis is aligned with the shear axis  $\hat{s}$ . Greater elongation (higher  $\xi_p$ ) in a given shear decreases the flipping frequency, which is found by integration of the angular velocity equation 1 to be  $\frac{\kappa(\xi_p + 1)}{\pi(\xi_p^2 + 2\xi_p + 2)}$ . At a given elongation  $\xi_p$ , higher shear also causes the polymer to rotate faster. To calculate the flipping motion of the polymer at each time step, we first project the polymer into the  $\hat{u}$ - $\hat{s}$  plane, calculate  $\phi$  and  $\xi_p$  from the projection, and then use the angular velocity equation 1 to calculate the rotation about the  $\hat{f}$  axis.

We now consider polymer deformations. During each computational time step, we assume the following:

1. The external streamwise strain rate  $\frac{\partial u}{\partial x}$  is constant over the length of the polymer during the step.
2. The external flow velocity is parallel to the flow axis  $\hat{u}$  over the length of the polymer during the step. Note that this assumption is necessarily true at the beginning of each step at the polymer center by construction of  $\hat{u}$ , and will remain approximately true for sufficiently small steps.
3. The polymer retains an equivalent elliptical shape, with no bending.

Forces acting on the polymer are the Stokes drag acting over the body of the polymer and the strain internal to the polymer. For simplicity, we now use a bead-and-spring dumbbell model of the polymer to calculate these forces. We model the polymer by two equivalent spheres of dimensional radius  $B\mathcal{H}$  joined by a spring of dimensional equilibrium length  $2B\mathcal{H}\xi$  which has no direct interaction with the flow. We adopt a variant of the Fraenkel-Warner spring (Bird, 1987) with dimensional restoring force  $\vec{F} = -\mathcal{G}BG(\xi)\hat{\xi}$ , where  $\mathcal{G}$  is a dimensional spring constant, and:

$$G(\xi) = (\xi - \xi_{eq}) \times \left\{ \begin{array}{ll} 1 & \text{if } \xi \leq \xi_{eq} \\ (1 + \xi - \xi_{eq})^{-p} & \text{if } \xi > \xi_{eq} \end{array} \right\} + \left\{ \begin{array}{ll} 0 & \text{if } \xi < \xi_W \xi_T \\ \frac{q(\xi - \xi_W \xi_T)^2}{(\xi - \xi_W)(2\xi_W \xi_T - \xi - \xi_W)} & \text{if } \xi \geq \xi_W \xi_T \end{array} \right\}. \quad (2)$$

Keyes and Abernathy (1987) have previously determined most of the parameter values for the above spring function by fitting the model to experiments. The equilibrium length of the polymer is set to  $\xi_{eq} = 2$  in order to match the observed frequency of polymer-induced fluctuations. The softening constant is set to  $p = 4$  to allow significant stretching to begin in the observed regime. The Warner constant is set to  $\xi_W = 100$  as representative of the maximum observed (or inferred) polymer elongation. We add the last term in the spring function equation 2, which activates at  $\xi_T = 0.9$  of the maximum extension with hardening constant  $q = 0.1$ , to model the eventual tightening of the polymer near its maximum extension. We choose this form so as to eliminate the sharp jerk which Keyes and Abernathy's (1987) polymer underwent when the polymer reached maximum extension. This function is graphed in Figure 2. While it possesses the qualitative behavior we seek in a relatively simple form, there are many other equally good alternatives. To account for hydrodynamic shielding of the two spheres at small separations, we multiply the Stokes drag by a factor  $\beta$  which is unity at infinite separation and approaches infinity as the spheres touch. This prevents interpenetration of the spheres. Brenner (1961) has calculated this  $\beta$  for Stokes flow, yielding:

$$\beta(\xi) = \frac{4}{3} \sinh \alpha \sum_{m=1}^{\infty} \frac{m(m+1)}{(2m-1)(2m+3)} \left[ \frac{4 \cosh^2(m + \frac{1}{2})\alpha + (2m+1)^2 \sinh^2 \alpha}{2 \sinh(2m+1)\alpha - (2m+1) \sinh 2\alpha} - 1 \right], \quad (3)$$

where  $\alpha \equiv \cosh^{-1} \xi$ . This function is plotted in Figure 3. Setting the sum of the Stokes and spring forces to zero yields a differential equation for the position of each of the spheres  $\vec{\mathcal{L}}_i$  in terms of dimensional variables:

$$\frac{d\vec{\mathcal{L}}_i}{dT} = \frac{\Delta\vec{U}_i}{\beta(\xi)} - \frac{\mathcal{G}G(\xi)}{6\pi\mu\beta(\xi)}\hat{\xi}, \quad (4)$$

where  $\Delta\vec{U}_i$  is the dimensional velocity difference between the ambient fluid at the  $i$ 'th sphere and at the polymer center. Since we are dealing with a flow that has nonzero strain rate in essentially only the streamwise direction, the dimensionless velocity difference is

$$\Delta\vec{U}_i = \frac{R}{R_\nu} [u_{\{0\}} + (x_i - x_0)u_x]\hat{x} + [v_{\{i\}} - v_{\{0\}}]\hat{y} + [w_{\{i\}} - w_{\{0\}}]\hat{z}, \quad (5)$$

where the subscripts  $i$  denote ambient quantities at the  $i$ 'th sphere, and where the subscripts zero denote quantities at the polymer center. When nondimensionalizing the position vector equation 4 with respect to the usual characteristic variables, we define  $C \equiv \frac{\mathcal{G}\mathcal{H}^2}{6\pi\mu B\nu R}$  as a dimensionless measure of the spring force versus the average viscous drag force on the spheres. This dimensionless spring



constant  $C$  is related to the analogous constant used in Keyes and Abernathy (1987) by  $C = C_{\text{Keyes}} \kappa$ , where  $\kappa$  is the dimensionless local shear.

Another issue to be addressed is whether the polymer should maintain constant volume during deformations, or whether it should maintain constant semi-minor axis, or whether it should possibly exhibit some intermediate behavior. This will be an important issue later when investigating how much deformation polymers undergo when embedded in a vortical flow. In the next section we will argue that the actual polymer behavior is probably approximately constant semi-minor axis at small deformations when polymer coils are still relatively tightly wound, and is approximately constant volume at large deformations when all coils have been effectively opened and the polymer has saturated with entrained fluid. For now, we will simply outline the calculation of the two extreme cases. In the constant semi-minor axis case, we just take the difference between the current positions of the two sphere centers to find the separation  $L$ , then calculate the current aspect ratio  $\xi = \frac{L}{2B}$ . In the constant volume case, the semi-minor axis length  $B$  is variable. We then combine the equation for the polymer volume  $V = \frac{4}{3}\pi(\xi + 1)B^3$  with the dimensionless deformation  $\xi = \frac{L}{2B} \equiv \frac{|\vec{L}_2 - \vec{L}_1|}{2B} \equiv \frac{|\vec{L}_2 - \vec{L}_1|}{2B\kappa}$  to eliminate  $B$ . This yields a cubic equation for the current deformation  $\xi$ . Therefore at the end of each computational step we use the known constant volume  $V$  and the new elongation  $L$  to solve for  $\xi$ , and then we solve for the new semi-minor axis  $B = \frac{L}{2\xi}$ . Since  $B$  is variable, the Stokes drag over the sphere decreases as its radius  $B$  decreases – followed to its logical conclusion, we would therefore decrease the dimensionless drag constant  $C$  as polymer extension  $\xi$  increases. Yet we expect the Stokes drag over the entire surface of the polymer to remain constant or perhaps even increase at larger polymer extensions. We model this expected behavior by retaining a constant dimensionless drag  $C$  at all polymer extensions. This drag constant  $C$  is taken to be that of the polymer at equilibrium ( $\xi = \xi_{eq}$ ). Therefore,  $C = \frac{6\eta^2}{\pi\mu\nu R} \sqrt{\frac{4\pi(\xi_{eq}+1)}{3V}}$ , where  $V$  is the given constant volume.

In summary, our computation for the motion of a polymer embedded in an external flow is:

- Construct local coordinate system:

$$\hat{u} = \frac{\vec{u}}{|\vec{u}|} = \frac{\langle u, v, w \rangle}{|\langle u, v, w \rangle|}, \quad (6)$$

$$\hat{s} = \frac{\nabla|\vec{u}| - \hat{u} \cdot \nabla|\vec{u}| \hat{u}}{|\nabla|\vec{u}| - \hat{u} \cdot \nabla|\vec{u}| \hat{u}|}, \quad (7)$$

$$\hat{f} = \hat{u} \times \hat{s}. \quad (8)$$

- Project polymer into  $\hat{u}$ - $\hat{s}$  plane. Determine  $\phi$  and  $\xi_p$  from the projection.
- Calculate rotation about  $\hat{f}$  axis:

$$\frac{d\phi}{dt} = \kappa \frac{(\xi_p + 1)^2 \cos^2 \phi + \sin^2 \phi}{(\xi_p + 1)^2 + 1}. \quad (9)$$

- Calculate  $\vec{L}_i = \langle x_i, y_i, z_i \rangle$  for both polymer ends ( $i = 1, 2$ ) due to deformation:

$$\frac{dx_i}{dt} = \frac{u_{\{i\}} + (x_i - x_0)u_x}{\beta(\xi)} \frac{R}{R_\nu} - C \frac{G(\xi)B}{\beta(\xi)} \frac{R}{R_\nu} (\hat{\xi}_i \cdot \hat{x}), \quad (10a)$$

$$\frac{dy_i}{dt} = \frac{v_{\{i\}} - v_{\{0\}}}{\beta(\xi)} - C \frac{G(\xi)B}{\beta(\xi)} \frac{R}{R_\nu} (\hat{\xi}_i \cdot \hat{y}), \quad (10b)$$

$$\frac{dz_i}{dt} = \frac{w_{\{i\}} - w_{\{0\}}}{\beta(\xi)} - C \frac{G(\xi)B}{\beta(\xi)} \frac{R}{R_\nu} (\hat{\xi}_i \cdot \hat{z}). \quad (10c)$$

- Superpose rotation and deformation to calculate new polymer position.
- If constant volume: calculate  $\xi$ , and then  $B$ , from:

$$\frac{6V}{\pi L^3} \xi^3 - \xi - 1 = 0, \quad (11)$$

$$B = \frac{L}{2\xi}. \quad (12)$$

- If constant semi-minor axis: calculate  $\xi = \frac{L}{2B}$ .

We embed the above polymer in the external flow of a streamwise vortex in a shear flow by specifying initial values for the position of the two spherical ends and for  $\xi$  (which in turn determines the initial semi-minor axis), and then running the flow model. After each computational flow step, we use the local flow velocities to calculate an updated shape, position, and orientation for the polymer.

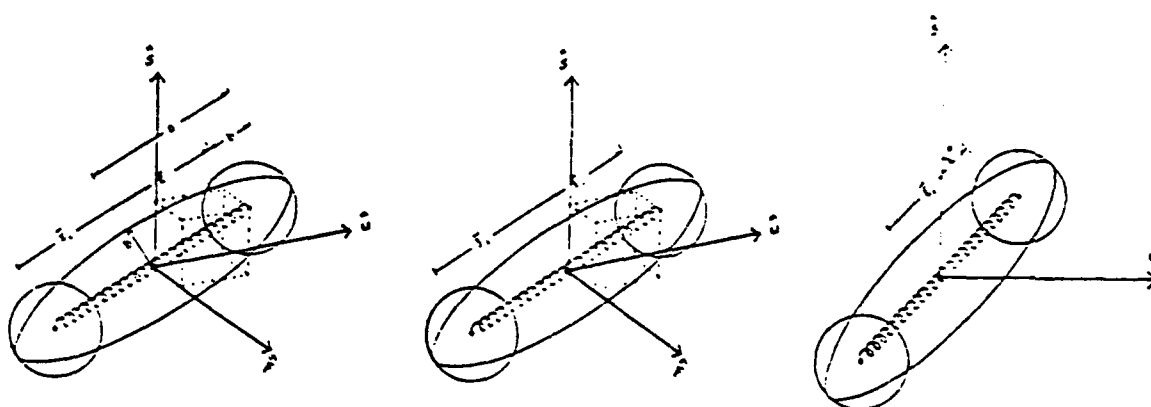


Figure 1: The hybrid polymer model, regarded as an ellipsoid for calculating rotation, and as a bead-and-spring dumbbell for calculating deformation. The first view shows dimensional lengths, the second view shows dimensionless lengths, and the third view shows dimensionless lengths projected into the  $u-s$  plane.

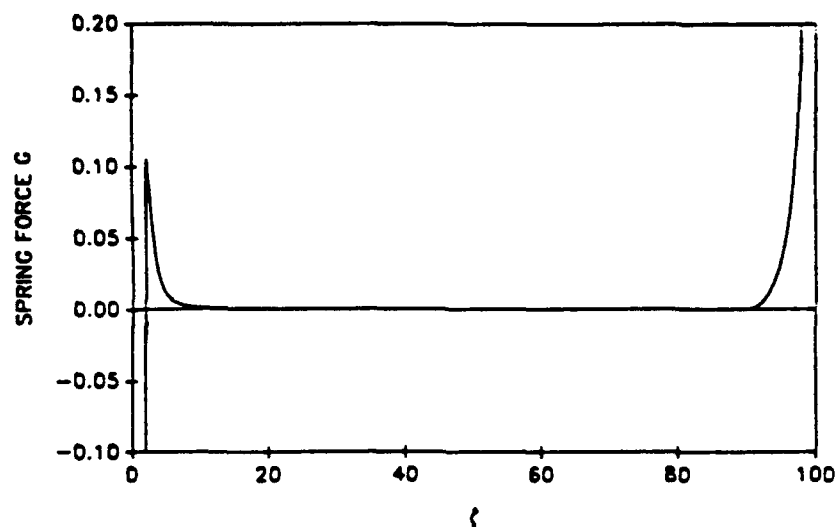


Figure 2: The dimensionless spring force  $G$  as a function of polymer extension  $\xi$ , as given by Equation 2. Note the repulsive force when compressed from equilibrium, and the smooth tightening as the spring approaches its maximum extension.

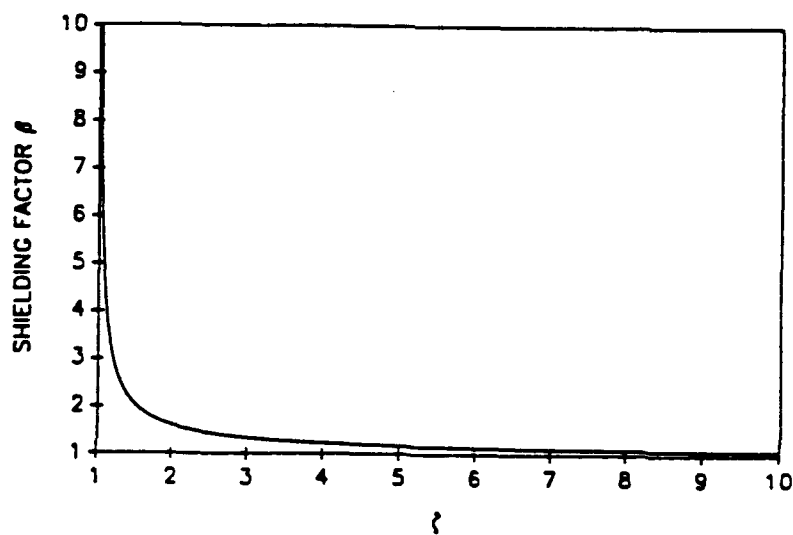


Figure 3: The Brenner correction factor  $\beta$  as a function of polymer extension  $\xi$ , as given by Equation 3. Multiplying the Stokes drag by this factor represents shielding of the radial fluid force on one sphere due to the presence of the other, and prevents interpenetration of the two spheres at small separations.

### 3 Benchmark: Comparison to Previous Results

In this section, we validate the above methodology for polymer dynamics by comparison with previous results. Since our model is a generalization of the model of Abernathy et al (1980) and Keyes and Abernathy (1987), we first replicate their results with our model, then perturb the flow slightly so as to still obtain essentially the same results, but by using some of the other logical branches in our numerical implementation.

We first consider a polymer molecule with the original spring constant<sup>1</sup> from Abernathy et al (1980) and Keyes and Abernathy (1987):

$$G(\xi) = \frac{\xi - \xi_{eq}}{1 - (\frac{\xi}{\xi_W})^2} \times \begin{cases} 1 & \text{if } \xi \leq \xi_{eq} \\ (1 + \xi - \xi_{eq})^{-p} & \text{if } \xi > \xi_{eq} \end{cases}. \quad (13)$$

Fixed parameters of the polymer are a maximum extension of  $\xi_W = 10$ , softening constant  $p = 0$ , and shielding function  $\beta(\xi) \equiv 1$ . We model the polymer as having constant semi-minor axis  $B$ ; in other words, as having variable volume. We embed this polymer in a uniform shear flow which can be described dimensionally as  $\vec{U} = \langle K\gamma, 0, 0 \rangle$ . This is the situation investigated by Abernathy et al (1980). For nondeformable ( $C \rightarrow \infty$ ) polymers whose equilibrium shapes are ellipsoids of various extensions  $\xi_{eq}$ , we generate rotation angles  $\phi$  versus time as shown in Figure 4. These curves correspond to Figure 6 of Abernathy et al (1980). Note that the flipping frequency decreases with increasing equilibrium extension  $\xi_{eq}$ . For polymers of various equilibrium shapes  $\xi_{eq}$  and of various stiffnesses  $C$ , we generate the flipping periods shown in Table 1. These correspond to selected points from Figure 8 of Abernathy et al (1980). In additional trials, we note the relative insensitivity of polymer extension to the maximum allowable extension  $\xi_W$ , as long as the allowable maximum remains large compared to the actual maximum deformation which the polymer undergoes. For polymers of equilibrium shape  $\xi_{eq} = 2$ , maximum extension  $\xi_W = 10$ , and of various stiffnesses  $C$ , we generate deformations  $\xi$  versus rotation angle  $\phi$  as shown in Figure 5. These correspond to Figure 9 of Abernathy et al (1980). In all cases, agreement is excellent.

Next, we again consider a polymer molecule with constant semi-minor axis and with spring constant given by Equation 13. Now we fix the polymer's parameters at an equilibrium extension of  $\xi_{eq} = 2$ , maximum extension  $\xi_W = 100$ , softening constant  $p = 4$ , and shielding function  $\beta(\xi)$  as given by Equation 3. For polymers of various stiffnesses  $C$ , we find the limit cycles of deformation versus rotation angle as shown in Figure 6. This figure corresponds to Figure 8 of Keyes and Abernathy (1987). Again, agreement is good. Early discrepancies in the analysis of this case revealed a typographical error in Equation 10b of Keyes and Abernathy (1987).<sup>2</sup>

In order to test other branches of our logic, we again consider a polymer with the same fixed parameters as above, and with stiffness  $C = 5$ , in flows slightly perturbed from that used above. The flows we test are listed in Table 2. In all cases, the results are practically the same as obtained above. This provides some degree of assurance that our various logical branches, used for dealing with different flow situations, are working as intended.

Finally, we test our logic for dealing with the streamwise strain rate component of a flow by embedding a nonresisting ( $C = 0$ ) polymer in a pure streamwise rate of strain flow  $\vec{u} = \langle 0.01(x + 0.005), 0, 0 \rangle$ . Although compressible, this flow is a simple test of our handling of the streamwise

<sup>1</sup>This original spring constant imparts a very sharp change in tension to the polymer as it nears its maximum extension  $\xi = \xi_W$ , much like a rope that has been snapped to its greatest extension. This was understandably of little concern to Keyes and Abernathy, given their modest polymer extensions in pure shear flow. Since we shall soon see that our flow will deform polymers to a much greater degree, we have replaced this original form with the more smoothly tightening form of Equation 2.

<sup>2</sup>The corrected equation reads:  $\dot{\xi} = \xi \sin \phi \cos \phi / \beta(\xi) - Cg(\xi) / \beta(\xi)$ .

Spring Stiffness ( $C$ )	Equilibrium Extension ( $\xi_{eq}$ )	Maximum Extension ( $\xi_{max}$ )	Period (to flip 180°)
0.01	0	0.00001	6.2819
0.01	1	1.87	9.7366
0.01	2	4.07	14.9224
0.1	2	3.97	14.7952
1	2	2.89	11.1434
10	2	2.10	10.4803
100	0	0.000005	6.2819
100	1	1.005	7.8541
100	2	2.01	10.4721

Table 1: Maximum extensions and flipping periods for polymers of various equilibrium shapes and stiffnesses in uniform shear flow. Other parameters are the same as for the polymers of Figure 4. Replication of selected points from Figure 8, Abernathy et al (1980).

Flow Type	Vortex Strength $R_v$	Local Velocity $u$	Streamwise Strain Rate $\frac{\partial u}{\partial x}$
$x$ independent	0	0.5	0
$x$ independent	0.1	0.5	0
$t$ independent	0.1	0.5	0 (held constant)
$t$ independent	0.1	0.5	$10^{-7}$ (held constant)
$t$ independent	0	0.75	0

Table 2: Slightly perturbed flow conditions for the polymer ( $C = 5$ ) of Figure 6. We duplicate that figure in each of these flows, testing other branches of our logic for different flow situations. Local shear is  $\frac{\partial u}{\partial y} = 1$  in all cases.

strain rate component of a 3-D flow. Results are in exact agreement with the analytic solution – the length of the polymer grows exponentially in time, and linearly with the streamwise position of the polymer center.

Although Keyes experimented with constant volume polymers, he does not report on the change in limit cycle obtained when changing from constant semi-minor axis to constant volume. For spring constant  $C = 5$  in uniform shear flow, we observe the maximum polymer deformation increase from  $\xi_{max} = 4.2$  to  $\xi_{max} = 18.1$ . This seems an unreasonably large extension in pure shear flow. We suspect that the actual polymer behavior is approximately constant semi-minor axis at small deformations when polymer coils are still relatively tightly wound, and is approximately constant volume at large deformations when all coils have been effectively opened and the polymer has saturated with entrained fluid. It therefore seems reasonable to treat polymers which undergo only modest deformations ( $\xi_{max} < 10$ ) under the constant semi-minor axis assumption as being well-described by that assumption, and to treat polymers which would undergo significant stretching ( $\xi_{max} > 10$ ) under the same assumption as actually being better modeled by the constant volume assumption. We implement this idea in our polymer simulations of the following section.

Now that we have tested and validated the polymer model, we are ready to check its predictions in a more interesting flow.

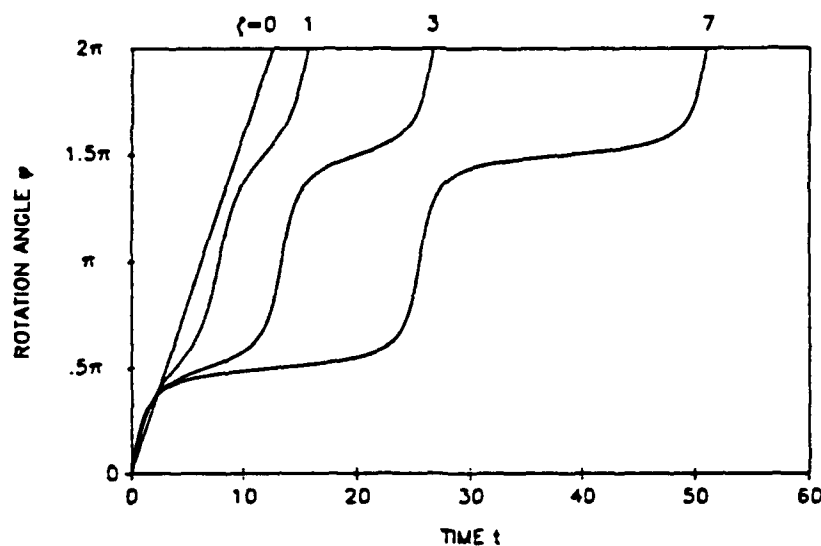


Figure 4: Rotation angle  $\phi$  versus time for nondeformable ( $C \rightarrow \infty$ ) polymers of equilibrium extension  $\xi_{eq} = 1, 2, 4, 8$  in uniform shear flow. Equation 13 describes the spring force, shielding function is  $\beta(\xi) \equiv 1$ , maximum extension is  $\xi_W = 10$ , and softening constant is  $p = 0$ . Replication of Figure 6, Abernathy et al (1980).

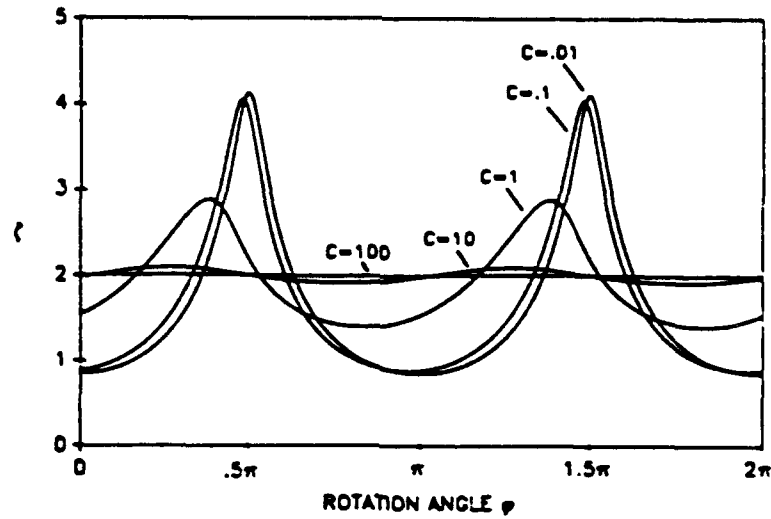


Figure 5: Polymer deformation  $\xi$  versus rotation angle  $\phi$  for polymers of stiffness  $C = 0.01, 0.1, 1, 10, 100$  in uniform shear flow. Equilibrium shape is  $\xi_{eq} = 2$ , and other parameters are the same as for the polymers of Figure 4. Replication of Figure 9, Abernathy et al (1980).

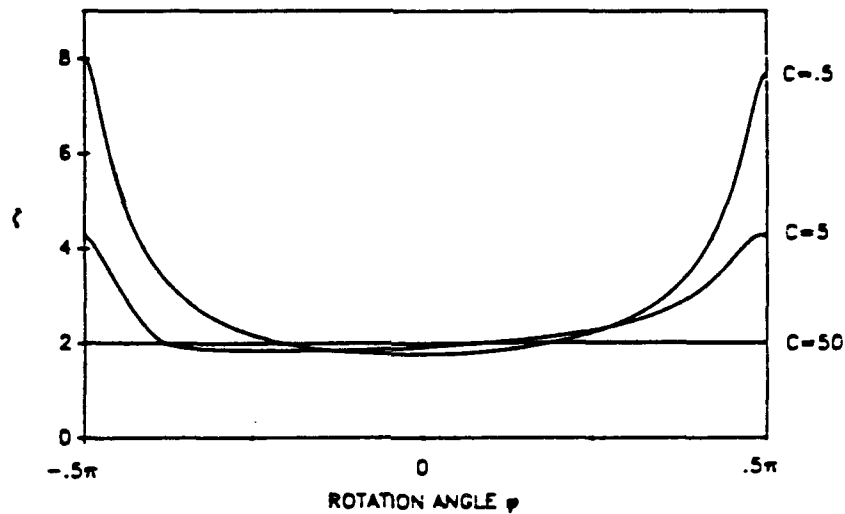


Figure 6: Limit cycle of polymer deformation  $\xi$  versus rotation angle  $\phi$  for polymers of stiffness  $C = 0.5, 5, 50$ . Equation 13 describes the spring force, shielding function  $\beta(\xi)$  is given by Equation 3, equilibrium shape is  $\xi_{eq} = 2$ , maximum extension is  $\xi_w = 100$ , and softening constant is  $p = 4$ . Replication of Figure 8, Keyes and Abernathy (1987).

## 4 Dynamics of the Embedded Polymer

We now investigate the dynamics of the polymer of Section 2 embedded in a streamwise independent vortical flow of a transient, streamwise independent flow and in a streamwise dependent flow, both of which are comprised of a diffusing streamwise vortex aligned in the flow direction of a shear flow. Our aim here is to determine whether such flow fields are capable of deforming embedded polymers significantly more than uniform shear flows.

We first show that significant polymer deformations do not occur in vortical flows which lack streamwise velocity gradients. We consider both streamwise independent vortical flows and streamwise dependent flows in which, within the numerical code, the local streamwise strain rate  $\frac{\partial u}{\partial x}$  in the vicinity of the polymer has been artificially set to zero. In both of these flows, a polymer of variable volume (constant semi-minor axis) flips and deforms in the same manner found by Keyes (and as verified in our benchmarks, Section 3) in uniform shear, as determined by the local shear in the vicinity of the polymer. This occurs regardless of initial location and initial orientation of the polymer, and regardless of vortex strength. This is due to the large difference between the characteristic length scales of the polymer and the vortex; on the length scale of the polymer, the local flow field due to the vortex is indistinguishable from a uniform shear flow. The only effect of the vortex is to rotate the local coordinate system within which the polymer executes its flipping motion. Peterlin (1970) anticipated this failure of the vortex crossflow to act as anything other than a new shear on the polymer's length scale, and tried to resolve the problem by proposing the polymers to be activated by the crossflow shear of "microvortices", whose crossflow length scale is comparable to that of the polymer. These microvortices would have to be nearly potential, and the polymer would have to be very close to the vortex center, for significant polymer deformation to occur. However, the number density of both polymers and of microvortices of the requisite size, strength, and concentration are low enough to make the required vortex-polymer pairing too rare an event to enable this to be a viable mechanism. We therefore conclude that significant deformations of a constant semi-minor axis polymer do not occur in vortical flows which lack streamwise velocity gradients, regardless of initial polymer location and orientation and regardless of the vortex strength.

We now consider a streamwise dependent vortical flow with streamwise velocity gradients. Since we find that a polymer embedded in such a flow experiences large deformations under the constant semi-minor axis assumption, we conclude that the polymer is probably better described as being of constant volume, as discussed at the end of the previous section. Figure 7 shows the deformation of a constant volume polymer near an initially-potential vortex of strength  $R_v = 6.5$  as the polymer is convected in the streamwise direction. Note that during the few flips before the streamwise gradient induces supercritical stretching (at streamwise positions  $x < 12$ ), the maximum polymer deformation is about a factor of four smaller than that expected in the same flow without a vortex. This is because the edge of the vortex core quickly diffuses past the polymer, thereby flattening the local shear near the polymer. At streamwise position  $x \approx 12$ , we see that the streamwise velocity gradient has grabbed the polymer and induced supercritical stretching. Figure 8 shows the deformation of a constant volume polymer near the core edge of a fully-diffused vortex (meaning near the flow boundary) of strength  $R_v = 6.5$ . Again, we see that the streamwise velocity gradient grabs the polymer during each flip cycle and induces significant stretching, with maximum deformation similar to that found near an initially-potential vortex. Experimentation shows that the threshold vortex strength required to induce these large extensions is  $R_v \approx 3$  near a potential vortex, or  $R_v \approx 5$  near the edge of a diffuse vortex. The streamwise velocity gradient grabs the polymer when it is at favorable  $\xi$ - $\phi$  orientations; the onset of large deformations can be relatively sensitive to this initial orientation, but polymers whose semi-major axis  $\hat{\xi}$  is initially within about  $25^\circ$  of the  $\hat{u}$  -  $\hat{s}$  plane achieve the required orientation within about one to three flips (Figure 9). Polymers that begin at



orientations outside of this range perform "log rolls" about the flipping axis  $\hat{f}$ , and do not experience significant stretching (Figure 10). So for a dilute concentration of polymers with initially random orientation located near such vortices, we can expect about 28% to be activated by the vortex. This yields an  $O(1)$  correction to the effective polymer concentration.

A positive streamwise velocity gradient is necessary to overcome the initial spring resistance. Thereafter, the streamwise gradient is relatively unimportant to further stretching, and a positive streamwise velocity difference between the two polymer ends is sufficient for continued stretching. This is demonstrated by the fact that once significant stretching has begun, a positive streamwise velocity difference can continue the stretching, even if the polymer passes through a region of negative streamwise velocity gradient. Polymers usually pass several times from regions with one sign of streamwise gradient to regions with the other sign of streamwise gradient. This is because there are either two or four such regions of alternating sign in the neighborhood of the vortex, and since polymers remain on roughly a cylindrical surface as they are convected around by the crossflow and downstream by the mean flow, therefore the crossflow naturally convects polymers around between the regions of streamwise velocity gradient with alternating sign.

Since the rate of spring softening at larger extensions (reflected in the softening constant  $p$  of the polymer model) was not as sharply determined as the other model constants in the original formulation by Keyes and Abernathy, it is appropriate to determine the sensitivity of our observed supercritical polymer stretching behavior to changes in spring softening. We find that qualitatively similar stretching occurs for springs with softening constant  $p > 2.7$  (Figure 11). This threshold corresponds to a maximum spring constant of  $G_{max} = 0.17$ , occurring at extension  $\xi = 2.6$ . Larger softening constants (such as Keyes and Abernathy's best fit of  $p = 4$ , which yields a maximum spring constant of  $G_{max} = 0.10$  at  $\xi = 2.3$ ) allow the positive streamwise velocity gradient to stretch the polymer more quickly, but to approximately the same extensions. Smaller softening constants  $p < 2.7$  correspond to greater maximum spring constants  $G_{max} > 0.17$ , which are strong enough to inhibit large stretching of the polymer at the streamwise strain rates typically found in our vortical flows.

Figures 7 and 8 illustrate how large deformations (on the order of  $\xi \sim 50$  to 60) can occur for vortices with different initial distributions of vorticity. This is one order of magnitude larger than the deformation obtained in laminar shear flow. Since the streamwise strain rate  $\frac{\partial u}{\partial x}$  is the physical mechanism responsible for initiating supercritical stretching of the polymers, we infer that large polymer deformations begin most often at the edge of the vortex core, where the streamwise strain rate  $\frac{\partial u}{\partial x}$  is a maximum. This statement holds on average, but not always in individual cases, since the onset of large polymer deformations is dependent to some degree on the initial orientation of the polymer.

Figure 12 shows how polymers at several different initial locations in the flow can all be subject to a large amount of deformation. Even a polymer that begins on the side of the vortex with negative streamwise strain rate is greatly elongated as it is convected around the vortex and into a region with positive strain rate.

Figure 13 shows how the maximum deformation of a polymer varies as a function of the vortex Reynolds number  $R_v$ . At each  $R_v$ , the maximum deformation shown is taken from over a variety of initial locations and orientations of the polymer. It appears that stronger vortices generally deform the polymers to a greater degree. Note that the polymer is inextensible past the Warner limit  $\xi_w = 100$ .

Constants in our adopted polymer model were fit by Keyes and Abernathy in accordance with data that for the most part involved polymers at small deformations. A degree of uncertainty in these parameters enters when we begin to deal with highly-deformed polymers. In fact, small changes in the polymer model at large deformations can significantly change the maximum polymer extension.

For example, Figure 14 shows the deformation of the polymers of Figures 7 and 8 under the change that the flipping rate  $\frac{d\phi}{dt}$  is reduced by half after a polymer extension of  $\xi > 10$ . In both cases, the polymer deforms to its maximum extensional limit. Thus, at these large polymer deformations, there is a degree of model uncertainty. Indeed, as large as our current polymer deformations are, they may be even larger with, for example, decreased flipping rate, increased spring softening, or increased Stokes drag with polymer extension.

The large polymer extensions demonstrated above raise the issue of polymer entanglement. Our non-bending polymer assumption seems quite reasonable for polymers undergoing stretching, but seems improbable during contraction, especially for highly elongated polymers. Memory and bending of highly elongated polymers during contraction make it seem likely that polymers take longer to collapse than to stretch, and that the polymer would undergo significant bending and buckling during this time. The flipping of these bent polymers suggests the possibility of polymer entanglements above some threshold concentration. In order to make a crude estimate as to this required concentration, assume a cubic lattice of polymers aligned in the flow direction with uniform separation  $S$ . The distances between polymer centers in the two crossflow directions are both  $2B + S$ , and the distance between polymer centers in the flow direction is  $2(\xi_{eq} + 1)B + S$ . Therefore the number density of polymers is  $\frac{1}{(2B+S)^2[2(\xi_{eq}+1)B+S]}$  and the volume concentration of polymers is

$\frac{(\xi_{eq}+1)B^3}{(2B+S)^2[2(\xi_{eq}+1)B+S]}$ . At an equilibrium extension of  $\xi_{eq} = 2$  and assuming that the polymer bends

approximately in half as it flips ( $S \sim \frac{\xi_{max}}{2} \approx 30B$ ), we anticipate significant polymer entanglements at volume concentrations greater than  $\sim 8 \times 10^{-5}$ . This agrees well with observed volume concentrations of polymer additives required for the onset of turbulence suppression and turbulent drag reduction. This equispaced cubic arrangement is a rather crude assumption; a more closely packed arrangement (such as hexagonal or face-centered cubic) seems more likely. Bertschy (1979) discusses how these other arrangements would change the above threshold volume concentration by a factor of at most  $\frac{71}{63}$ , an increase of only 13%. It should be noted that with the onset of polymer entanglements at large extensions, the validity of our polymer model in this regime is again suspect.

In this section, we have presented evidence of a heretofore unnoticed mechanism for the activation of polymers in a turbulent flow. This mechanism is the generation in vortical flows of streamwise strain rates of sufficient strength to stretch nearby polymers to large elongations. With the resumption of flipping by these highly deformed polymers, we have also demonstrated a viable mechanism for intermolecular entanglements in dilute polymer solutions. In the next section, we infer how these elongated polymers affect the evolving flow field.

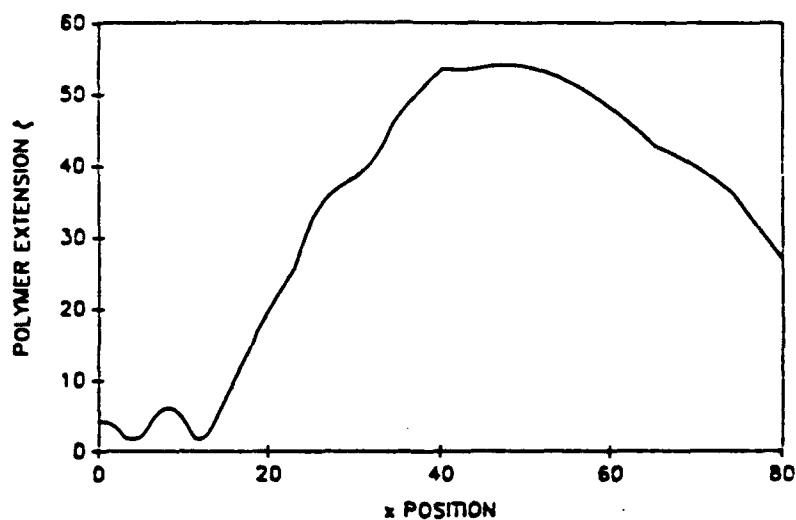


Figure 7: Extension of a polymer located near an initially-potential streamwise vortex with streamwise velocity gradients. The streamwise gradient soon overcomes the maximum spring force and begins supercritical stretching of the polymer. Relaxation of the polymer is eventually caused by rotation of the polymer axis  $\xi$  past the flow axis  $\hat{u}$ .  $C = 5$ ,  $a = 2$ ,  $R_\nu = 6.5$ ,  $R = 3000$ , polymer is initially located at  $z = -0.1$ ,  $y = 0.5$ ,  $x = 0$  and oriented with its semi-major axis parallel to the  $\hat{x}$  axis.

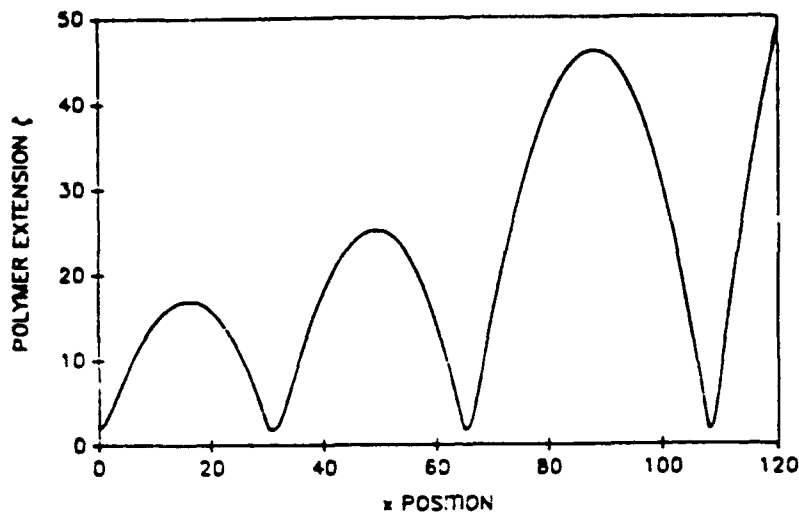


Figure 8: Extension of a polymer located near the core edge (flow boundary) of a fully diffused stream-wise vortex with streamwise gradients. The streamwise velocity gradient again induces significant polymer stretching, with maximum deformations similar to those found near an initially-potential vortex.  $C = 5$ ,  $a = 2$ ,  $R_v = 6.5$ ,  $R = 3000$ , polymer is initially located at  $z = -0.9$ ,  $y = 0.5$ ,  $x = 0$ , and is initially oriented with its semi-major axis parallel to the  $\hat{y}$  axis.

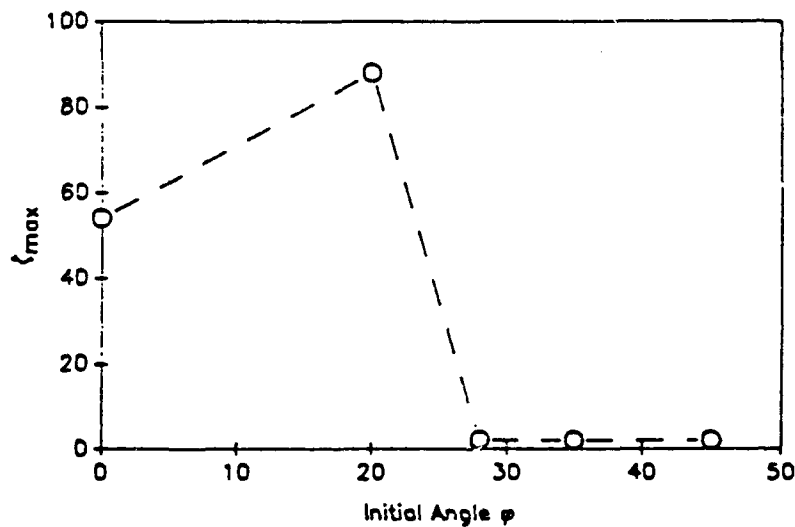


Figure 9: Maximum deformation of the polymer of Figure 7 when started with semi-major axis in the  $\hat{x} - \hat{z}$  plane and at various angles from the  $\hat{x}$  axis. Large deformations occur when the polymer is within about  $25^\circ$  of the  $\hat{x}$  axis.

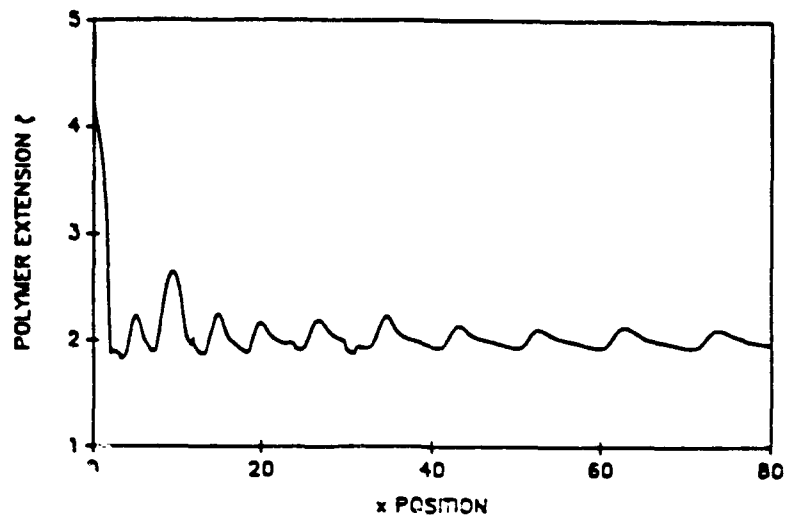


Figure 10: Extension of a polymer located near an initially-potential streamwise vortex with streamwise velocity gradients. The polymer is initially located in the  $\hat{x}-\hat{z}$  plane with a  $35^\circ$  angle between its semi-major axis and  $\hat{x}$ . The polymer's semi-major axis becomes generally aligned with the flipping axis  $\hat{f}$  and performs "log rolls".  $C = 5$ ,  $a = 2$ ,  $R_v = 6.5$ ,  $R = 3000$ , polymer is initially located at  $z = -0.1$ ,  $y = 0.5$ ,  $x = 0$ .

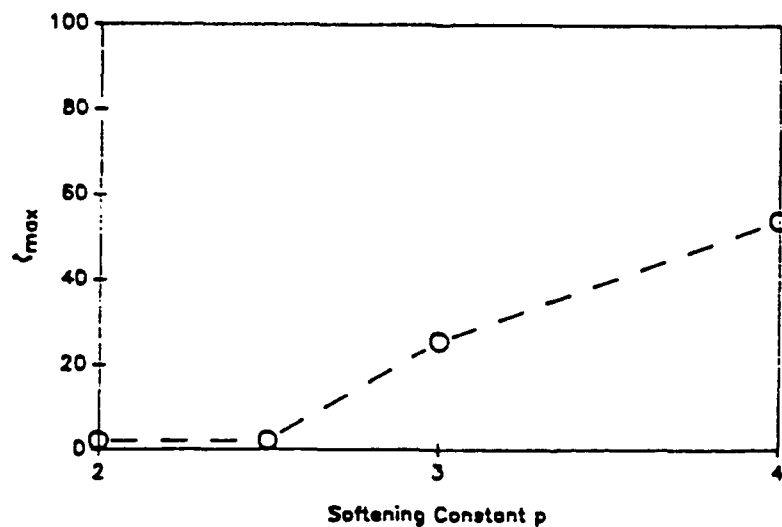


Figure 11: Maximum deformation of the polymer of Figure 7 for various values of the spring softening constant  $p$ . Large deformations occur for values greater than about 2.7.

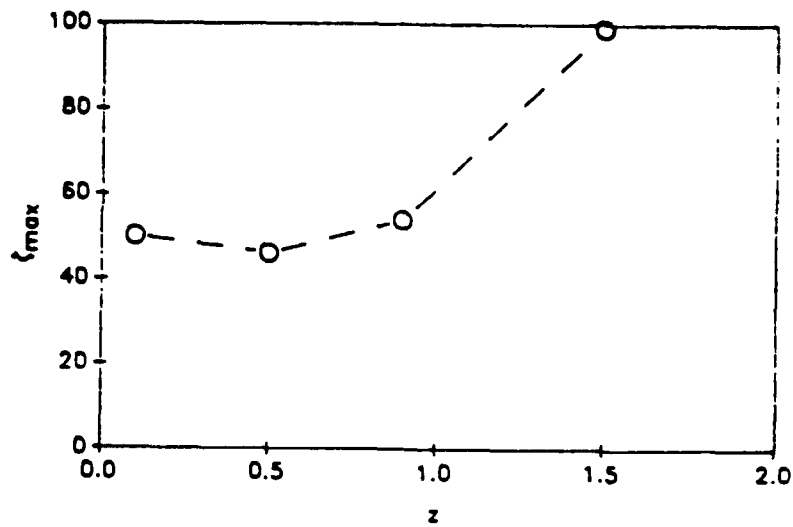


Figure 12: Maximum deformation of the polymer of Figure 7 for various initial spanwise locations. Large deformations can occur for a wide range of initial spanwise locations. Even polymers that begin on the side of the vortex with negative streamwise strain rate are expanded as they are convected into the positive strain rate region.

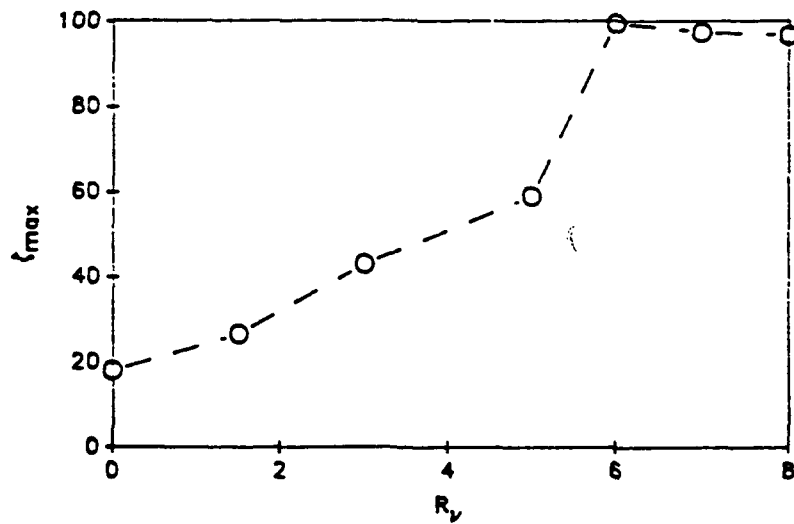


Figure 13: Maximum deformation of a polymer near vortices of various vortex Reynolds numbers. Stronger vortices seem to induce larger maximum deformations, until the Warner limit ( $\xi_w = 100$ ) is reached.  $C = 5$ ,  $a = 2$ ,  $R = 3000$ . The maximum deformation is taken over all polymers whose initial orientations are as in Figure 9 and whose initial locations are as in Figure 12.

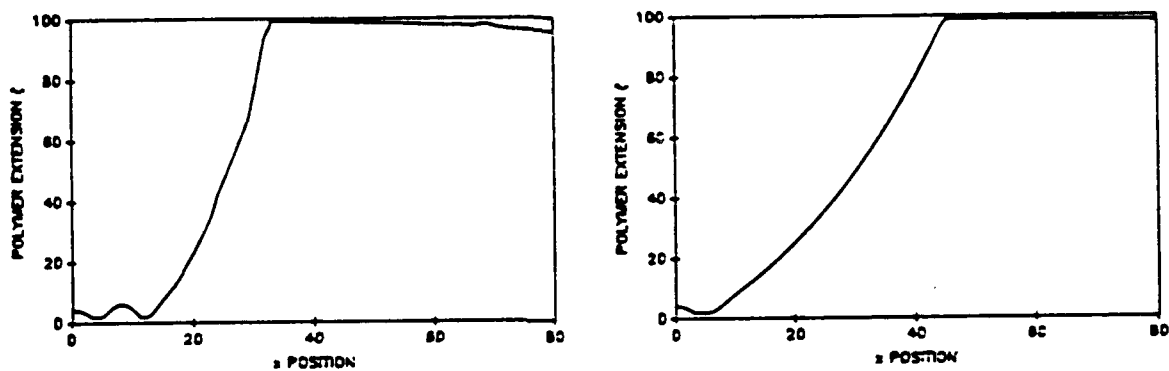


Figure 14: Deformation of the polymers of Figures 7 and 8 under the change that the flipping rate is halved after the polymer is significantly stretched. Deformation ceases after the polymer is almost fully extended. This demonstrates how small changes in the polymer model can cause significant changes in the dynamics of highly deformed polymers.

## 5 Implications for Flow Stability

The development of highly elongated polymers in a streamwise dependent vortical flow implies an increase in the local viscosity due to increased energy dissipation in the form of a secondary flow around the polymers. We now address several related questions in a relatively nonrigorous way.

Where does an increase in the flow viscosity occur? The development of elongated polymers begins in regions with positive streamwise strain rate, but then stretching can continue into other flow regions with positive streamwise velocity difference between the polymer ends, even if the streamwise strain rate is negative.<sup>3</sup> Therefore polymer extension is very memory-dependent, and cannot be readily predicted from local flow quantities. We do not do so here, but since polymers maintain a relatively constant distance from the vortex center and since stretched polymers persist over some time, it may be reasonable to predict a maximum polymer extension at all points in the flow from some local flow quantity, then take the maximum extension over all angles at a given radius from the vortex center and attribute that extension to all points at the given radius. Polymer stretching is more easily initiated when the streamwise strain rate  $\frac{\partial u}{\partial x}$  is high, so it appears that supercritical stretching of the polymers begins most readily at the edge of the vortex core, where the streamwise strain rate is highest. Therefore it seems likely that the local increase in viscosity occurs mainly at the edge of and within the vortex core, as the moving core edge will have swept over and activated polymers in that region as it diffuses outward.

What is the local increase in flow viscosity due to the extended polymer? Rheological investigations on this general topic seem to fall into two classes: those in pure shear flow, and those in pure elongational flow. In pure shear flow, Einstein (1906) and Jeffery (1922) investigated the secondary flow around a dilute solution of rigid spheres and rigid ellipsoids, respectively, to determine the local increase in flow viscosity. Eirich (1956) discusses the extension of these results to a concentrated solution of rigid spheres as a function of their volume fraction. Keyes (1987) approximates the viscosity increase due to a dilute suspension of deforming ellipsoids as the sum of rotational and deformational energy dissipation components. Graessley (1965) investigates a deforming entanglement network and demonstrates shear-thinning, but makes no connection between solvent viscosity and the zero-shear viscosity of his entangled network solution. In pure elongational flow, Peterlin (1966) has studied dilute dumbbell and necklace models with their more numerous degrees of freedom, and King and James (1983) have investigated similar models which are "frozen" after partial elongation due to intramolecular entanglements. Takserman-Krozer and Ziabicki (1963) have studied the energy dissipation of dilute rigid ellipsoids. Leal and Hinch (1973) have been concerned with dilute particles under the influence of Brownian couples in both shearing and strain rate flows. In a much-quoted work, Batchelor (1971) investigated the additional stress generated by both a dilute solution and by a hydrodynamically concentrated solution<sup>4</sup> of rigid rods. Ryskin (1987) incorporated Batchelor's results into his view of an uncoiling ("yo-yo") polymer which always has a taut central section of some effective length. This notion appears useful in that it explains the experimental observation that stressed polymers nearly always break in the middle. Also, Batchelor's treatment of a hydrodynamically concentrated rigid rod solution seems in rough agreement with the more recent tube model of De Gennes (1971) and of Doi and Edwards (1986), which views hydrodynamically concentrated polymers as each confined to a tube with streamwise orientation. The enclosed polymer moves (reptates) along the length of the tube on a long time scale and moves along

<sup>3</sup> Recall from Section 4.3 that the vortex creates a region of positive streamwise strain rate on one side and a region of negative streamwise strain rate on the other side. Therefore the vortex crossflow naturally convects an embedded polymer from one region to the other.

<sup>4</sup> A hydrodynamically concentrated solution is one in which there are significant dynamical interactions between adjacent particles, even though the net volume concentration of solute may remain quite low.



the short diameter of the tube, making contact with neighboring tubes, on a much shorter time scale. The Ryskin/Batchelor viewpoint appears sufficiently consistent with this current thought on near-neighbor interactions in a hydrodynamically concentrated solution and seems applicable to our situation, therefore we follow their argument for the local increase in flow viscosity. As in Ryskin (1987), we begin with an equivalent form of Batchelor's (1971) expression for the viscosity increase due to a suspension of hydrodynamically concentrated rigid rods in a strain rate flow:

$$\frac{\nu_1}{\nu_0} = 1 + \frac{\pi N(\xi B)^3}{9 \ln \frac{\pi}{\Phi}}, \quad (14)$$

where  $\nu_0$  is the kinematic viscosity of the solvent,  $\nu_1$  is the kinematic viscosity of the solution,  $N$  is the number density of polymers per volume, and  $\Phi$  is the hydrodynamically effective volume concentration based on a cycle-averaged polymer deformation. Ryskin neglects the dependence of volume concentration  $\Phi$  on extension  $\xi$  (as we do for highly deformed polymers in our constant-polymer-volume assumption), and argues that increases in viscosity are significant only when adjacent polymers are hydrodynamically interactive. The hydrodynamically effective volume concentration of polymers can be expressed as:

$$\Phi = \frac{4}{3} \pi \left( \frac{\xi_{cs} B}{3.6} \right)^3 N, \quad (15)$$

where  $\xi_{cs}$  is the maximum stable elongation of the polymer before coil-stretch transition of the polymer begins, marking the onset of supercritical stretching. The length  $\frac{\xi_{cs} B}{3.6}$  is the hydrodynamically effective radius of the unstretched polymer, an expression derived by Rabin et al (1985). This effective concentration can also be approximated using Einstein's result:

$$\Phi = \frac{2}{5} C [\eta], \quad (16)$$

where  $C$  is the concentration of polymers by weight and  $[\eta]$  is the intrinsic viscosity.<sup>5</sup> Substituting the effective concentration equations 15 and 16 into the viscosity equation 14 yields an expression for the increase in viscosity:

$$\frac{\nu_1}{\nu_0} = 1 + \frac{1.555 C [\eta]}{\ln \frac{2.5\pi}{C[\eta]}} \left( \frac{\xi_{max \text{ effective}}}{\xi_{cs}} \right)^3, \quad (17)$$

where  $\xi_{max \text{ effective}}$  is the effective maximum elongation (in a cycle-averaged sense) of the polymer in the flow. From our polymer model, the maximum stable elongation of a polymer of spring constant  $C = 5$  in a streamwise strain rate flow is roughly  $\xi_{cs} \sim 4$ . We assume that the effective maximum elongation in a strain rate flow is half the maximum extension when flipping and entangling; this is roughly  $\frac{\xi_{max}}{4}$ , or  $\frac{60}{4}$  (from Figures 7 and 8). Typical polymer concentration and intrinsic viscosity values, such as in the sink flow experiment of James and Saringer (1980), yield:

$$C [\eta] \approx \left( 0.00002 \frac{g}{cm^3} \right) \left( 2500 \frac{cm^3}{g} \right) \quad (18a)$$

$$\approx 0.05 \quad (18b)$$

$$\frac{\nu_1}{\nu_0} \approx 1 + \frac{1.555 \cdot 0.05}{\ln \frac{2.5\pi}{0.05}} \left( \frac{60}{4} \right)^3 \quad (18c)$$

$$\approx 1.8. \quad (18d)$$

<sup>5</sup> Intrinsic viscosity is defined as  $[\eta] \equiv \lim_{C \rightarrow 0} \frac{\eta_1 - \eta_0}{C\eta_0}$ , where  $\eta_0$  denotes the shear viscosity of the solvent and  $\eta_1$  denotes the shear viscosity of the solution.

Thus it seems that for the deformations  $\xi_{max} \sim 60$  that we observe for polymers embedded in a typical streamwise dependent vortical flow field, the local viscosity may be almost doubled.

What effect does the increase in local viscosity have on the velocity profiles? The rigorous way to address this question would be to adapt our fluid equations of motion to account for variable viscosity during the flow-evolution calculations. We do not pursue this because it is impractical to fill our computational domain with polymer molecules, and because the significant memory effects of the polymer seem to thwart attempts to characterize the local deformation of polymers (and thus the local increase in viscosity) by using local flow quantities. Therefore our first order answer as to the effect on the velocity profiles is to say that, since the prominent energy dissipation mechanisms of polymers are likely to be active both as they are being significantly stretched and as they are interacting when collapsing and buckling during the flipping cycle, therefore the viscosity will be approximately doubled, per our above calculation, over the core of the vortex (since polymers seem to be activated principally at the edge of the vortex core, and remain active for some period thereafter in regions over which the core edge has already swept). However, we must be a little cautious here. If the local viscosity is doubled, then the local vortex Reynolds number  $R_\nu$  is halved, decreasing the streamwise strain rate, thereby decreasing the deformation of embedded polymers, thereby decreasing the local viscosity, etc. Obviously, a rigorous treatment would account for this feedback mechanism; omitting the feedback from calculations means that our observed polymer deformations and inferred viscosity increase can only be considered an upper bound to the true rheological effect. As an extreme case, one could even guess that the tiniest increase in viscosity decreases the streamwise strain rate enough so that supercritical stretching of polymers is completely halted, meaning that the polymer additives have practically no effect on the flow. However, this extreme suggestion is ruled out by our earlier sensitivity analysis. Recall from Section 4 that for our typical initially-potential vortices, nearby polymers undergo about the same magnitude of supercritical stretching for vortices of strength from  $R_\nu > 3$  up to at least 6.5. Therefore concentrated vortices of strength  $R_\nu = 6.5$  (among the most destabilizing vortices found in a boundary layer flow) can double the viscosity due to polymer activation with effects that are generally undiminished by this feedback mechanism. However, similar vortices which begin at weaker strengths, or more diffuse vortices whose threshold strength seems to be on the order of  $R_\nu \approx 5$  (both of which are less destabilizing in a boundary layer flow), may induce effects that are limited – to some extent not determined here – by feedback. Therefore, it appears that in streamwise vortical flows, polymer additives can reduce the vortex Reynolds number of at least the most destabilizing vortices by half. This significantly reduces the inflectional profiles within the vortex core.

We now have a fairly complete chain of evidence for turbulence suppression and turbulent drag reduction by dilute polymer additives. We are led to surmise that, in accordance with the stability calculations of Pearson (1985) and in analogy to inviscid stability theory, the inflectional velocity profiles at the vortex center associated with vortices of strength  $R_\nu \sim 6.5$  are unstable. The flow in the vicinity of a vortex experiences a local increase in viscosity due to polymer deformation. We infer from this the local decrease in strength of the associated vortex (to  $R_\nu \sim 3$ ) via polymer deformation, and thus the local decrease in strength of the associated vortex, the reduced inflectional velocity profiles within the vortex core, and the increased stability of the flow within the region of the vortex.

The author gratefully acknowledges the support provided by the Academic Research Directorate, US Military Academy, and by the Army Research Laboratory for providing computing resources for this work.

## Bibliography

1. Abernathy, F., Bertschy, J., Chin, R., and Keyes, D. (1980). Polymer-induced fluctuations in high-strain-rate laminar flows. *J. Rheology* **24**, 647-665.
2. Acierno, D., Titomanlio, G., and Nicodemo, L. (1974). Elongational flow of dilute polymer solutions. *Rheologica Acta* **13**, 1040-1044.
3. Batchelor, G. (1970). Slender-body theory for particles of arbitrary cross-section in Stokes flow. *J. Fluid Mech.* **44**, 419-440.
4. Batchelor, G. (1971). The stress generated in a non-dilute suspension of elongated particles by pure straining motion. *J. Fluid Mech.* **46**, 813-829.
5. Bertschy, J. (1979). Laminar and turbulent boundary layer flows of drag reducing solutions. Ph.D. Thesis, Division of Applied Sciences, Harvard University.
6. Bird, R., Hassager, R., Armstrong, R., and Curtiss, C. (1987). *Dynamics of Polymeric Liquids*, 2d Ed., Vol II. Wiley, New York.
7. Brenner, H. (1961). The slow motion of a sphere through a viscous fluid towards a plane surface. *Chem. Eng. Sci.* **16**, 242-251.
8. Cottrell, F., Merrill, E., and Smith, K. (1970). Intrinsic viscosity and axial extension ratio of random-coiling macromolecules in a hydrodynamic shear field. *J. Polymer Sci. A-2*, **8**, 287.
9. De Gennes, P. (1971). Reptation of a polymer chain in the presence of fixed obstacles. *J. Chem. Phys.* **55**, 2, 572.
10. Doi, M., and Edwards, S. (1986). *The Theory of Polymer Dynamics*. Clarendon, Oxford.
11. Einstein, A. (1906). A new determination of molecular dimensions. *Annalen der Physik* **19**, 289 (and **34**, 591).
12. Eirich, F. (1956). *Rheology*, Vol I. Academic Press, New York.
13. Graessley, W. (1965). Molecular entanglement theory of flow behavior in amorphous polymers. *J. Chem. Phys.* **43**, 8, 2696-2703.
14. Hinch, E. (1977). Mechanical models of dilute polymer solutions in strong flows. *Physics of Fluids* **20**, S22-S30.
15. James, D., and Saringer, J. (1980). Extensional flow of dilute polymer solutions. *J. Fluid Mech.* **97**, 655-671.
16. Jeffery, G. (1922). The motion of ellipsoidal particles immersed in a viscous fluid. *Proc. Royal Soc. London A* **102**, 161-179.
17. Keyes, D., and Abernathy, F. (1987). A model for the dynamics of polymers in laminar shear flows. *J. Fluid Mech.* **185**, 503-522.
18. King, D., and James, D. (1983). Analysis of the Rouse model in extensional flow. II. Stresses generated in sink flow by flexible macromolecules and by finitely extended macromolecules. *J. Chem. Phys.* **78**, 4749-4754.

19. Leal, L., and Hinch, E. (1973). Theoretical studies of rigid particles affected by Brownian couples. *Rheologica Acta* **12**, 127-132.
20. Myers, J., and Abernathy, F. (1989). Characteristics and stability implications of a stream-wise vortex in bounded shear flow, *Transactions of the Seventh Army Conference on Applied Mathematics and Computing*, 371-427.
21. Pearson, C. (1985). A class of instabilities associated with streamwise vorticity. Ph.D. Thesis, Division of Applied Sciences, Harvard University.
22. Pearson, C., and Abernathy, F. (1984). Evolution of the flow field associated with a streamwise diffusing vortex. *J. Fluid Mech.* **146**, 271-283.
23. Peterlin, A. (1966). Hydrodynamics of linear macromolecules. *Pure & Applied Chem.* **12**, 563-586.
- Peterlin, A. (1970). Molecular model of drag reduction by polymer solutes. *Nature* **227**, 598-599.
25. Rabin, Y., Henyey, F., and Pathria, R. (1985). Theoretical studies of the coil stretching transition of polymers in elongational flows. In *Polymer-Flow Interaction, AIP Conf. Proc.*, Vol 137 (ed. Y. Rabin), 43-58.
26. Roache, P. (1982). *Computational Fluid Dynamics*. Hermosa, Albuquerque.
27. Rubin, R., and Mazur, J. (1975). Ordered spans of unrestricted and self-avoiding random-walk models of polymer chains. I. Space-fixed axes. *J. Chem. Phys.* **63**, 5362-5374.
28. Ryskin, G. (1987a). Calculation of the effect of polymer additive in a converging flow. *J. Fluid Mech.* **178**, 423-440.
29. Ryskin, G. (1987b). Turbulent drag reduction by polymers: a quantitative theory. *Phys. Rev. Letters* **59**, 2059-2062.
30. Suri, A. (1988). Streamwise vortices in shear flow transition. Ph.D. Thesis, Division of Applied Sciences, Harvard University.
31. Takserman-Krozer, R., and Ziabicki, A. (1963). Behavior of polymer solutions in a velocity field with parallel gradient. *J. Polymer Sci. A*, **1**, 491-515.
32. Toms, B. (1948). Some observations on the flow of linear polymer solutions through straight tubes at large Reynolds numbers. *Proc. 1st Int. Congr. Rheol.* **2**, 135-141.
33. Yang, Z. (1987). A single streamwise vortical structure and its instability in shear flows. Ph.D. Thesis, Division of Applied Sciences, Harvard University.

# A MASSIVELY PARALLEL ITERATIVE NUMERICAL ALGORITHM FOR IMMISCIBLE FLOW IN NATURALLY FRACTURED RESERVOIRS

Jim Douglas, Jr.\*    P. J. Paes Leme<sup>†</sup>    Felipe Pereira<sup>‡</sup>  
Li-Ming Yeh<sup>‡</sup>

**Abstract.** We propose a new iterative numerical scheme designed for massively parallel processing for an immiscible displacement in a naturally fractured reservoir. The procedure is based on a domain decomposition technique applied to a mixed finite element approximation of the problem; the domain is decomposed into individual elements. Numerical experiments are presented to illustrate its performance on a CM-5 system.

## 1 INTRODUCTION

High quality numerical simulations of fluid flow in petroleum reservoirs require the use of increasingly finer grids in the numerical discretization of the governing system of partial differential equations so that a large number of length scales relevant to the problem can be incorporated into the simulations. This problem is critical when inhomogeneities are present and their influence need to be adequately resolved.

Detailed two-dimensional studies of the effect of the inhomogeneities of a single porosity medium have reached the limits of existing serial computers [22], [23]. For dual porosity models, meaningful studies are infeasible on serial machines.

In this paper, a parallel iterative procedure, specially designed for massively parallel processing, is proposed for the numerical solution of dual-porosity models for immiscible flow in a naturally fractured reservoir. Two implementations were performed, one completely portable, adequate for MIMD systems, and the other using a data-parallel programming language particular to a Connection Machine Model CM-5 using SIMD control. For fluid flow simulations with grid sizes relevant for applications we find a good scalability of our algorithm, with a consistent slightly better performance of the MIMD version. However, through the use of vector processing units, the SIMD code runs faster.

The model problem treated in this paper corresponds physically to a waterflooding of a naturally fractured petroleum reservoir where the average spacing between fractures is relatively small compared to the reservoir size. With the terminology adopted in previous works the particular model treated herein is known as the "medium block model" [16]. The system of partial differential equations governing fluid flow in this double porosity formulation treats the flow in each matrix block in a completely parallelizable fashion. The part of the system describing the flow in the fractures is then

---

\*Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395

<sup>†</sup>Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, Brazil

<sup>‡</sup>Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395

Work supported in part by the Army Research Office contract number DAAL03-89-C-0038  
with the University of Minnesota Army High Performance Computing Research Center

numerically approximated by a hybridized mixed finite element method. A domain decomposition technique in which the domain is decomposed into individual elements is then applied. This allows us to adapt the solution of the problem to massively parallel processing. Domain decomposition techniques distinct from the one used here can be found in [7], [24], [25], [19], and [18]. As our first step towards full three-dimensional fluid flow simulations, we consider in the numerical studies described in this work two-dimensional fractured media to which are attached three-dimensional matrix blocks.

We define a nonlinear iterative procedure and use it to solve numerically the part of the system describing flow in the fractures which is coupled to the system of equations for the matrix blocks through source terms. This method is motivated by the linear problem analyzed in [17], which is closely related to the one introduced in [9] for a Helmholtz problem and extended to another Helmholtz-like problem related to Maxwell's equations [8], [10]. As in the above references, we shall make use of the hybridization of mixed finite element methods introduced in [21] and [20] more than twenty-five years ago and which has been carefully analyzed in [1]; see also [4], [2], and [3]. For the numerical solution of the local problems associated with the matrix blocks, a simple finite difference scheme [16] will be used. A rigorous proof of convergence of the iterative procedure is currently being investigated by the authors. For the simpler problem of two-phase flow through a single porosity medium, convergence of the iteration has been established.

This paper is organized in the following way. In §2 a brief description of the model considered here, along with a time discretization for it, is given. A domain decomposition technique and the new iterative procedure defined for a mixed finite element approximation of the system of equations in the fractures appear in §3. A detailed description of the time-dependent algorithm developed to solve the full governing system appears in §4. Distinct parallel implementations of our numerical method in a CM-5 system are analyzed in §5. Finally, §6 is devoted to our conclusions and interesting open problems related to this work.

## 2 THE MODEL PROBLEM

### 2.1 Governing Equations

We consider saturated, two-phase, incompressible, immiscible flow, the phases being  $o$  (oil or nonwetting phase) and  $w$  (water or wetting phase), with densities and viscosities  $\rho_\alpha$  and  $\mu_\alpha$ ,  $\alpha = o, w$ , respectively. See [15] in this volume for a description of the system of equations governing fluid flow in a single porosity model under the above assumptions.

The governing system for the medium block model is derived through the mathematical theory of homogenization [12]. It produces a two-phase, single-porosity model for the flow in the matrix system and a second, slightly modified single-porosity system in the fractures. To reduce the number of subscripts in the notation, we use capital letters to indicate quantities in the fractures and small letters to indicate those in the matrix blocks.

Let  $\Omega_x$  denote the block attached to the point  $x \in \Omega$ ; the  $w$ -saturation in  $\Omega_x$  will be indicated by  $s(x, y, t)$ ,  $x \in \Omega$ ,  $y \in \Omega_x$ ,  $t \geq 0$ , etc.

The capillary pressure and relative permeability functions are somewhat different

in the fractures than in the matrix blocks. Generally, one assumes that the fractures are essentially like spaces between two parallel planes and that  $S_{\min} = 0$  and  $S_{\max} = 1$ . The singularity in the capillary pressure curve as  $S$  decreases to  $S_{\min} = 0$  is weaker than that for the capillary pressure function in the blocks, and the relative permeability functions can be taken to be linear or nearly linear. The absolute permeability tensor on the fracture sheet reflects the geometry of the blocks [11].

The source terms in the saturation and pressure equations in the fractures contain two terms, one defining the external flow (wells in practice). In addition, there are matrix source terms  $q_{m,\alpha}$ ,  $\alpha = o, w$ , for each of the phases. The system governing flow in the fracture system can be written as

$$\Phi \frac{\partial S}{\partial t} + \nabla \cdot Q_w = q_{\text{ext},w} + q_{m,w} \quad \text{for } x \in \Omega, \quad t > 0, \quad (2.1a)$$

$$Q_w = -\tilde{\Lambda}_w(S) \nabla \Psi_w \quad \text{for } x \in \Omega, \quad t > 0, \quad (2.1b)$$

$$-\Phi \frac{\partial S}{\partial t} + \nabla \cdot Q_o = q_{\text{ext},o} + q_{m,o} \quad \text{for } x \in \Omega, \quad t > 0, \quad (2.1c)$$

$$Q_o = -\tilde{\Lambda}_o(S) \nabla \Psi_o \quad \text{for } x \in \Omega, \quad t > 0, \quad (2.1d)$$

$$S = P_c^{-1}(\Psi_c + (\rho_o - \rho_w)gz), \quad (2.1e)$$

Incompressibility requires that  $q_{m,o} + q_{m,w} = 0$ .

It is convenient to write the equations on the block  $\Omega_x$  as

$$\phi \frac{\partial s}{\partial t} - \nabla \cdot [\tilde{\lambda}_w(s) \nabla \psi_w] = 0 \quad \text{for } y \in \Omega_x, \quad t > 0, \quad (2.2a)$$

$$-\nabla \cdot [\tilde{\lambda}(s) \nabla \psi_w + \tilde{\lambda}_o(s) \nabla \psi_c] = 0 \quad \text{for } y \in \Omega_x, \quad t > 0, \quad (2.2b)$$

$$s = p_c^{-1}(\psi_c + (\rho_o - \rho_w)gz); \quad (2.2c)$$

we have assumed that the external sources affect the fracture system only. The boundary conditions for the matrix problems are given by requiring continuity of the potentials:

$$\psi_w(x, y, t) = \Psi_w(x, t) \quad \text{for } y \in \partial\Omega_x, \quad x \in \Omega, \quad t > 0, \quad (2.3a)$$

and

$$\psi_c(x, y, t) = \Psi_c(x, t) \quad \text{for } y \in \partial\Omega_x, \quad x \in \Omega, \quad t > 0. \quad (2.3b)$$

The matrix source terms are defined as follows. The volume of the  $w$ -fluid leaving the block  $\Omega_x$  is

$$\int_{\partial\Omega_x} v_w \cdot n \, da(y) = \int_{\Omega_x} \nabla \cdot v_w \, dy = - \int_{\Omega_x} \phi \frac{\partial s}{\partial t} \, dy;$$

consequently, let

$$q_{m,w}(x, t) = -\frac{1}{|\Omega_x|} \int_{\Omega_x} \phi \frac{\partial s}{\partial t} \, dy \quad \text{for } x \in \Omega, \quad t > 0. \quad (2.4)$$

We complete the model by specifying the external boundary conditions and the initial conditions for the system. For the case of no flow across the external boundary,

$$\tilde{\Lambda}_\alpha(s) \nabla \Psi_\alpha \cdot n = 0 \quad \text{for } x \in \partial\Omega, \quad t > 0, \quad \alpha = o, w. \quad (2.5)$$

Initial saturations (i.e., capillary potentials) must be specified:

$$\begin{aligned}\Psi_c(x, 0) &= \Psi_{\text{init},c}(x) \quad \text{for } x \in \Omega, \\ \psi_c(x, y, 0) &= \psi_{\text{init},c}(x, y) \quad \text{for } y \in \Omega_x, x \in \Omega.\end{aligned}$$

To be consistent, (2.3) and (2.5) should hold when  $t = 0$ .

## 2.2 Time-Discretization

Discretize the time variable by choosing  $t^0, t^1, t^2, \dots, t^N$  such that  $0 = t^0 < t^1 < t^2 < \dots < t^N$ , and set  $\Delta t^n = t^n - t^{n-1}$ . An approximation to a function  $\Theta$  related to the fracture system at a point  $x \in \Omega$  at time  $t^n$  will be denoted by

$$\Theta^n \approx \Theta(x, t^n).$$

Approximate (2.1) implicitly by backwards Euler approximations in time to obtain the system

$$\Phi \frac{S^n - S^{n-1}}{\Delta t^n} + \nabla \cdot Q_w^n = q_{\text{ext},w}^n + q_{m,w}^n \quad \text{for } x \in \Omega, t > 0, \quad (2.6a)$$

$$Q_w^n = -\tilde{\Lambda}_w(S^n) \nabla \Psi_w^n \quad \text{for } x \in \Omega, t > 0, \quad (2.6b)$$

$$-\Phi \frac{S^n - S^{n-1}}{\Delta t^n} + \nabla \cdot Q_o^n = q_{\text{ext},o}^n + q_{m,o}^n \quad \text{for } x \in \Omega, t > 0, \quad (2.6c)$$

$$Q_o^n = -\tilde{\Lambda}_o(S^n) \nabla \Psi_o^n \quad \text{for } x \in \Omega, t > 0, \quad (2.6d)$$

$$S^n = P_c^{-1}(\Psi_c^n + (\rho_o - \rho_w)gz), \quad (2.6e)$$

The time-discretization of the equations describing the flow in the matrix blocks will be discussed in the context of a finite difference discretization of the matrix system in §4.

## 3 DOMAIN DECOMPOSITION FOR THE FRACTURE SYSTEM

Parallelization of the solution of the global fracture system problem is achieved through a spatial decomposition, which we now describe.

Let  $\Omega \subset R^d$ ,  $d = 2$  or  $3$ , be a bounded domain with a Lipschitz boundary  $\partial\Omega$ . Let  $\{\Omega_j, j = 1, \dots, M\}$  be a partition of  $\Omega$ :

$$\bar{\Omega} = \cup_{j=1}^M \bar{\Omega}_j; \quad \Omega_j \cap \Omega_k = \emptyset, \quad j \neq k.$$

Assume that  $\partial\Omega_j$ ,  $j = 1, \dots, M$ , is also Lipschitz and that  $\Omega_j$  is star-shaped. In practice, with the exception of perhaps a few  $\Omega_j$ 's along  $\partial\Omega$ , each  $\Omega_j$  would be convex with a piecewise-smooth boundary. Let

$$\Gamma = \partial\Omega, \quad \Gamma_j = \Gamma \cap \partial\Omega_j, \quad \Gamma_{jk} = \Gamma_{kj} = \partial\Omega_j \cap \partial\Omega_k.$$

Let us consider decomposing (2.6) over the partition  $\{\Omega_j\}$ . In addition to requiring  $\{S_j^n\}$ ,  $\{Q_{\alpha,j}^n\}$ ,  $\{\Psi_{\alpha,j}^n\}$ ,  $\alpha = w, o$ ,  $j = 1, \dots, M$ , to be a solution of (2.6) for  $x \in \Omega_j$ ,  $j = 1, \dots, M$  it is necessary to impose the consistency conditions

$$\Psi_{\alpha,j} = \Psi_{\alpha,k}, \quad x \in \Gamma_{jk}, \quad \alpha = w, o, \quad (3.1a)$$

$$Q_{\alpha,j} \cdot \nu_j + Q_{\alpha,k} \cdot \nu_k = 0, \quad x \in \Gamma_{jk}, \quad \alpha = w, o, \quad (3.1b)$$

where  $\nu_j$  is the unit outer normal to  $\Omega_j$ .



### 3.1 Weak Formulation

Let  $V_j = H(\text{div}, \Omega_j)$  and  $W_j = L^2(\Omega_j)$  for  $j = 1, \dots, M$ . The weak formulation of (2.6) with the domain decomposed according to the discussion above is given by seeking  $\{S_j^n, Q_{w,j}^n, Q_{o,j}^n, \Psi_{w,j}^n, \Psi_{o,j}^n\} \in W_j \times V_j \times V_j \times W_j \times W_j$ ,  $j = 1, \dots, M$ , such that

$$\begin{aligned} & \frac{(\Phi S_j^n, w_1)_{\Omega_j} - (\Phi S_j^{n-1}, w_1)_{\Omega_j}}{\Delta t^n} + (\nabla \cdot Q_{w,j}^n, w_1)_{\Omega_j} \\ & = (q_{\text{ext},w}^n, w_1)_{\Omega_j} + (q_{m,w}^n, w_1)_{\Omega_j}, \end{aligned} \quad (3.2a)$$

$$\left( \frac{Q_{w,j}^n}{\bar{\Lambda}_w(S_j^n)}, v_1 \right)_{\Omega_j} - (\Psi_{w,j}^n, \text{div } v_1)_{\Omega_j} + \langle \Psi_{w,j}^n, v_1 \cdot \nu \rangle_{\partial \Omega_j} = 0, \quad (3.2b)$$

$$\begin{aligned} & - \frac{(\Phi S_j^n, w_2)_{\Omega_j} - (\Phi S_j^{n-1}, w_2)_{\Omega_j}}{\Delta t^n} + (\nabla \cdot Q_{o,j}^n, w_2)_{\Omega_j} \\ & = (q_{\text{ext},o}^n, w_2)_{\Omega_j} + (q_{m,o}^n, w_2)_{\Omega_j}, \end{aligned} \quad (3.2c)$$

$$\left( \frac{Q_{o,j}^n}{\bar{\Lambda}_o(S_j^n)}, v_2 \right)_{\Omega_j} - (\Psi_{o,j}^n, \text{div } v_2)_{\Omega_j} + \langle \Psi_{o,j}^n, v_2 \cdot \nu \rangle_{\partial \Omega_j} = 0, \quad (3.2d)$$

$$(P_c(S_j^n), w_3)_{\Omega_j} = (\Psi_{c,j}^n, w_3)_{\Omega_j} + ((\rho_o - \rho_w)gz, w_3)_{\Omega_j}, \quad (3.2e)$$

where  $v_1, v_2 \in V_j$  and  $w_1, w_2, w_3 \in W_j$ . There is a technical difficulty with (3.2b) and (3.2d); the meaning of the restriction of an  $L^2$ -function on  $\Omega_k$  to  $\Gamma_{jk}$  is not clear. Thus, (3.2) is properly viewed as motivation for the treatment of the discrete case to be discussed below.

### 3.2 Mixed Finite Element Approximation

We shall treat the case in which  $\{\Omega_j\}$  is a partition of  $\Omega$  into individual elements (simplices, rectangles, prisms), though an inspection of the procedure would indicate that larger subdomains are permissible. Let  $W^h \times V^h$  be a mixed finite element space over  $\{\Omega_j\}$ ; any of the usual choices is acceptable: [4], [2], [3], [6], [26], [28], [29]. Each of these spaces is defined through local spaces  $V_j^h \times W_j^h = V(\Omega_j) \times W(\Omega_j)$ , and setting

$$\begin{aligned} V^h &= \{v \in H(\text{div}, \Omega) : v|_{\Omega_j} \in V_j^h\}, \\ W^h &= \{w : w|_{\Omega_j} \in W_j^h\}. \end{aligned}$$

In each space  $W^h$  in the various families of mixed elements referenced above, the functions  $w \in W^h$  are allowed to be discontinuous across each  $\Gamma_{jk}$ . As a consequence, attempting to impose the consistency conditions (3.1) would force a flux conservation error; i.e., (3.1b) would not be satisfied unless the approximate solution  $\Psi_\alpha^h \in W^h$ ,  $\alpha = w, o$ , to the discrete analogue of (3.1b) is constant, a totally uninteresting case. So, let us introduce Lagrange multipliers [21], [20], [1] on the edges  $\{\Gamma_{jk}\}$ . In the discussion below we consider the parameter  $\alpha$  to be either  $o$  or  $w$ . Assume that, when  $Q_{\alpha,j} = Q_{\alpha,j}^h|_{\Omega_j}$ ,  $Q_{\alpha,j}^h \in V^h$ , its normal component  $Q_{\alpha,j} \cdot \nu_j$  on  $\Gamma_{jk}$ , is a polynomial of some fixed degree  $\tau_\alpha$ , where for simplicity we shall assume  $\tau_\alpha$  independent of  $\Gamma_{jk}$  (see [5] if not). Set

$$\Lambda_\alpha^h = \{\lambda_\alpha : \lambda_\alpha|_{\Gamma_{jk}} \in P_{\tau_\alpha}(\Gamma_{jk}) = \Lambda_{\alpha,jk}, \Gamma_{jk} \neq \emptyset\};$$

note that there are two copies of  $P_{\tau_\alpha}$  assigned to the set  $\Gamma_{jk}$ :  $\Lambda_{\alpha,jk}$  and  $\Lambda_{\alpha,kj}$ .

The hybridized mixed finite element method is given by dropping the superscript  $h$  and seeking

$$\{S_j^n \in W_j, Q_{\alpha,j}^n \in V_j, \Psi_{\alpha,j}^n \in W_j, \lambda_{\alpha,jk}^n \in \Lambda_{\alpha,jk}\},$$

where  $j = 1, \dots, M; k = 1, \dots, M; \alpha = w, o$ , such that

$$\begin{aligned} & \frac{(\Phi S_j^n, w_1)_{\Omega_j} - (\Phi S_j^{n-1}, w_1)_{\Omega_j}}{\Delta t^n} + (\nabla \cdot Q_{w,j}^n, w_1)_{\Omega_j} \\ &= (q_{\text{ext},w}^n, w_1)_{\Omega_j} + (q_{m,w}^n, w_1)_{\Omega_j}, \end{aligned} \quad (3.3a)$$

$$\left( \frac{Q_{w,j}^n}{\bar{\Lambda}_w(S_j^n)}, v_1 \right)_{\Omega_j} - (\Psi_{w,j}^n, \text{div } v_1)_{\Omega_j} + \sum_k \langle \lambda_{w,jk}^n, v_1 \cdot \nu \rangle_{\Gamma_{jk}} = 0, \quad (3.3b)$$

$$\begin{aligned} & - \frac{(\Phi S_j^n, w_2)_{\Omega_j} - (\Phi S_j^{n-1}, w_2)_{\Omega_j}}{\Delta t^n} + (\nabla \cdot Q_{o,j}^n, w_2)_{\Omega_j} \\ &= (q_{\text{ext},o}^n, w_2)_{\Omega_j} + (q_{m,o}^n, w_2)_{\Omega_j}, \end{aligned} \quad (3.3c)$$

$$\left( \frac{Q_{o,j}^n}{\bar{\Lambda}_o(S_j^n)}, v_2 \right)_{\Omega_j} - (\Psi_{o,j}^n, \text{div } v_2)_{\Omega_j} + \sum_k \langle \lambda_{o,jk}^n, v_2 \cdot \nu \rangle_{\Gamma_{jk}} = 0, \quad (3.3d)$$

$$(P_c(S_j^n), w_3)_{\Omega_j} = (\Psi_{c,j}^n, w_3)_{\Omega_j} + ((\rho_o - \rho_w)gz, w_3)_{\Omega_j}, \quad (3.3e)$$

where  $v_1, v_2 \in V_j$  and  $w_1, w_2, w_3 \in W_j$ .

### 3.3 The Iterative Method

In order to define an iterative method for solving the above system [9], [8] it is convenient to replace (3.3b) and (3.3d) by the Robin transmission boundary condition

$$-\beta Q_{\alpha,j} \cdot \nu_j + \Psi_{\alpha,j} = \beta Q_{\alpha,k} \cdot \nu_k + \Psi_{\alpha,k}, \quad x \in \Gamma_{jk} \subset \partial\Omega_j, \alpha = w, o, \quad (3.4a)$$

$$-\beta Q_{\alpha,k} \cdot \nu_k + \Psi_{\alpha,k} = \beta Q_{\alpha,j} \cdot \nu_j + \Psi_{\alpha,j}, \quad x \in \Gamma_{kj} \subset \partial\Omega_k, \alpha = w, o, \quad (3.4b)$$

where  $\beta$  is a positive (normally chosen to be a constant) function on  $\cup \Gamma_{jk}$ .

Now we formulate an iterative version of the finite element approximation of (3.3) with consistency conditions given by (3.4). Consider the Lagrange multiplier to be  $\lambda_{\alpha,jk}$ ,  $\alpha = w, o$  as seen from  $\Omega_j$  and  $\lambda_{\alpha,kj}$ ,  $\alpha = w, o$  as seen from  $\Omega_k$ . Then, modify (3.4) to read

$$-\beta Q_{\alpha,j} \cdot \nu_j + \lambda_{\alpha,jk} = \beta Q_{\alpha,k} \cdot \nu_k + \lambda_{\alpha,kj}, \quad x \in \Gamma_{jk} \subset \partial\Omega_j, \alpha = w, o,$$

$$-\beta Q_{\alpha,k} \cdot \nu_k + \lambda_{\alpha,kj} = \beta Q_{\alpha,j} \cdot \nu_j + \lambda_{\alpha,jk}, \quad x \in \Gamma_{kj} \subset \partial\Omega_k, \alpha = w, o,$$

so that

$$\langle \lambda_{\alpha,jk}, v \cdot \nu_j \rangle_{\Gamma_{jk}} = \langle \beta(Q_{\alpha,j} \cdot \nu_j + Q_{\alpha,k} \cdot \nu_k) + \lambda_{\alpha,kj}, v \cdot \nu_j \rangle_{\Gamma_{jk}}, \alpha = w, o.$$

The objective of a domain decomposition iterative method is to localize the calculations to problems over smaller domains than  $\Omega$ . Here, it is feasible to localize to each  $\Omega_j$  by evaluating the quantities in (3.3) related to  $\Omega_j$  at the new iterate level and those in (3.3) related to neighboring subdomains  $\Omega_k$  such that  $\Gamma_{jk} \neq \emptyset$  at the previous iteration level. Specifically, the algorithm would be as follows: let, for all  $j$  and  $k$ ,

$$S_j^{n-1} \in W_j, Q_{\alpha,j}^{n-1} \in V_j, \Psi_{\alpha,j}^{n-1} \in W_j, \lambda_{\alpha,jk}^{n-1} \in \Lambda_{\alpha,jk}, \lambda_{\alpha,kj}^{n-1} \in \Lambda_{\alpha,kj}, \alpha = w, o,$$

( $\lambda_{jk}^0 = \lambda_{kj}^0$  seems natural) be the solution of the discretized system of equations at some discrete time (we introduce in the notation the superscript  $i$  which is an iteration counter).

Then the solution propagated by one time step is given as the limit as  $i \rightarrow \infty$  of recursive solutions of the equations

$$\frac{(\Phi S_j^{n,i}, w_1)_{\Omega_j} - (\Phi S_j^{n-1}, w_1)_{\Omega_j}}{\Delta t^n} + (\nabla \cdot Q_{w,j}^{n,i}, w_1)_{\Omega_j} \quad (3.5a)$$

$$= (q_{\text{ext},w}^{n,i}, w_1)_{\Omega_j} + (q_{m,w}^{n,i}, w_1)_{\Omega_j},$$

$$\left( \frac{Q_{w,j}^{n,i}}{\bar{\Lambda}_w(S_j^{n,i-1})}, v_1 \right)_{\Omega_j} - (\Psi_{w,j}^{n,i}, \text{div } v_1)_{\Omega_j} + \sum_k \langle \beta Q_{w,j}^{n,i} \cdot \nu_j, v_1 \cdot \nu_j \rangle_{\Gamma_{jk}} \quad (3.5b)$$

$$= - \sum_k \langle \beta Q_{w,k}^{n,i-1} \cdot \nu_k + \lambda_{w,kj}^{n,i-1}, v_1 \cdot \nu_j \rangle_{\Gamma_{jk}},$$

$$- \frac{(\Phi S_j^{n,i}, w_2)_{\Omega_j} - (\Phi S_j^{n-1}, w_2)_{\Omega_j}}{\Delta t^n} + (\nabla \cdot Q_{o,j}^{n,i}, w_2)_{\Omega_j} \quad (3.5c)$$

$$= (q_{\text{ext},o}^{n,i}, w_2)_{\Omega_j} + (q_{m,o}^{n,i}, w_2)_{\Omega_j},$$

$$\left( \frac{Q_{o,j}^{n,i}}{\bar{\Lambda}_o(S_j^{n,i-1})}, v_2 \right)_{\Omega_j} - (\Psi_{o,j}^{n,i}, \text{div } v_2)_{\Omega_j} + \sum_k \langle \beta Q_{o,j}^{n,i} \cdot \nu_j, v_2 \cdot \nu_j \rangle_{\Gamma_{jk}} \quad (3.5d)$$

$$= - \sum_k \langle \beta Q_{o,k}^{n,i-1} \cdot \nu_k + \lambda_{o,kj}^{n,i-1}, v_2 \cdot \nu_j \rangle_{\Gamma_{jk}}.$$

The Lagrange multipliers are updated according to

$$\lambda_{w,jk}^{n,i} = \beta(Q_{w,j}^{n,i} \cdot \nu_j + Q_{w,k}^{n,i-1} \cdot \nu_k) + \lambda_{w,kj}^{n,i-1}, \quad (3.5e)$$

$$\lambda_{o,jk}^{n,i} = \beta(Q_{o,j}^{n,i} \cdot \nu_j + Q_{o,k}^{n,i-1} \cdot \nu_k) + \lambda_{o,kj}^{n,i-1}, \quad (3.5f)$$

and finally the equation for the capillary pressure is linearized:

$$\left( \frac{\partial P_c(S_j^{n,i-1})}{\partial S} (S_j^{n,i} - S_j^{n,i-1}), w_3 \right)_{\Omega_j} \quad (3.5g)$$

$$= (\Psi_{o,j}^{n,i} - \Psi_{w,j}^{n,i}, w_3)_{\Omega_j} + ((\rho_o - \rho_w)gz, w_3)_{\Omega_j} - (P_c(S_j^{n,i-1}), w_3)_{\Omega_j}.$$

We still have to explain how the matrix source terms are incorporated into the iterative procedure. We will postpone this discussion to §4.

We have been able to prove the following theorem concerning the convergence of the iterative procedure defined above in the simplified context where matrix blocks are suppressed from the model.

**Theorem 3.1 (Convergence of the Iterative Procedure)** *Suppose that a smooth solution of the system (2.1) exists. Then, there exists a constant  $t^*$  such that, when  $\Delta t \leq t^*$ ,*

1) *the iterative scheme (3.5) converges; i.e., there exists*

$$\{S_j^n \in W_j, Q_{o,j}^n \in V_j, \Psi_{o,j}^n \in W_j, \lambda_{o,jk}^n \in \Lambda_{o,jk}\}$$

such that

$$\|S_j^{n,i} - S_j^n\| + \|Q_{\alpha,j}^{n,i} - Q_{\alpha,j}^n\| + \|\Psi_{\alpha,j}^{n,i} - \Psi_{\alpha,j}^n\| + \|\lambda_{\alpha,jk}^{n,i} - \lambda_{\alpha,jk}^n\| \rightarrow 0$$

as  $i \rightarrow \infty$ ; moreover,  $\lambda_{\alpha,jk}^n = \lambda_{\alpha,kj}^n$ ,  $\alpha = w, o$ ; and

2) the above limit converges to the smooth solution in the sense there is a constant  $c$  such that

$$\|S(t^n) - S^n\| + \|Q_\alpha(t^n) - Q_\alpha^n\| + \|\Psi_\alpha(t^n) - \Psi_\alpha^n\| \leq c(\Delta t + h),$$

where  $\Theta^n(x) = \Theta_j^n(x)$ , for  $x \in \Omega_j$  and  $h$  represents the partition of  $\Omega$ .

A rigorous proof of the above statement will appear elsewhere.

#### 4 THE COMPUTATIONAL ALGORITHM

Our numerical procedure will combine a computationally inexpensive finite difference procedure to solve the local problems associated with each matrix block with a hybridized mixed finite element method applied to the global fracture system problem. The fracture and matrix systems cannot be handled sequentially, since a small change in the boundary values on each matrix block can cause flow of a volume of fluid that is large in comparison to the volume of the fractures. The matrix-fracture interaction for the medium block model can be handled implicitly by a linearization of the matrix problems to be made precise below. The final procedure requires solution of many small linear systems, each corresponding to an element of the discretized fracture system and its associated matrix block. The solution of these small and uncoupled linear systems can be handled easily by a parallel machine.

Discretize the space variables by defining grids over  $\Omega$  and over each matrix block  $\Omega_x$ . We consider  $\Omega$  and  $\Omega_x$ ,  $x \in \Omega$ , to be rectangular parallelepipeds; more general domains can be treated by either finite difference or finite element techniques quite analogous to the methods to be described herein. Suppose that  $\Omega = [0, D_1] \times [0, D_2] \times [0, D_3]$ . Then, divide each  $D_j$  into  $N_j$  intervals, which for simplicity we take to be of equal size  $H_j = D_j/N_j$ ,  $j = 1, 2, 3$ . Thus, the set

$$\mathcal{G}_f = \{x_L = (L_1 H_1 + H_1/2, L_2 H_2 + H_2/2, L_3 H_3 + H_3/2) : L_j = 0, 1, \dots, N_j - 1; j = 1, 2, 3\},$$

consists of the centers of the elements of the mixed method.

Again for notational convenience assume that the matrix blocks are all of the same size and consider a grid defined on the matrix block  $\Omega_x$  which will be used in a finite difference discretization of the matrix equations to be discussed below. Let  $h_j$  and  $n_j$  be analogous to  $H_j$  and  $N_j$  and set

$$\mathcal{G}_m = \{y_\ell = (\ell_1 h_1, \ell_2 h_2, \ell_3 h_3) : \ell_j = 0, 1, 2, \dots, n_j, j = 1, 2, 3\}.$$

Also, let

$$\mathcal{I}_m = \{y_\ell = (\ell_1 h_1, \ell_2 h_2, \ell_3 h_3) : \ell_j = 1, 2, \dots, n_j - 1, j = 1, 2, 3\}$$

indicate the interior nodes and  $\partial\mathcal{G}_m = \mathcal{G}_m \setminus \mathcal{I}_m$  the boundary nodes. (Advantage should be taken of any symmetry of the solution on a matrix block to allow the solution to be computed only at necessary nodes.)

An approximation to a function  $\Theta$  related to the the fracture system at a point  $x_L \in \mathcal{G}_f$  will be denoted by

$$\Theta_L^n \approx \Theta(x_L, t^n).$$

and for a function  $\theta$  associated with the block at the point  $x_L \in \mathcal{G}_f$ , denote the approximation to  $\theta$  at  $y_\ell \in \mathcal{G}_m$  by

$$\theta_{L,\ell}^n \approx \theta(x_{L,\ell}, t^n),$$

where  $x_{L,\ell} = x_L - y_\ell$  (this places a top corner of the block at  $x_L$ ).

The matrix equations will be completely linearized, but *not* the fracture equations. The discrete matrix system is directly solvable. The four parts of the algorithm below uncouple the calculations related to the matrix blocks from those of the fracture calculation:

i) Initialization. For each  $L$  and  $\ell$ , set

$$\begin{aligned} \Psi_{c,L}^0 &= \Psi_{\text{init},c}(x_L), & S_L^0 &= P_c^{-1} \left( \Psi_{c,L}^0 + (\rho_o - \rho_w) g z(x_L) \right), \\ \psi_{c,L,\ell}^0 &= \psi_{\text{init},c}(x_{L,\ell}), & s_{L,\ell}^0 &= p_c^{-1} \left( \psi_{c,L,\ell}^0 + (\rho_o - \rho_w) g z(x_{L,\ell}) \right). \end{aligned}$$

ii) Matrix system. For each  $L$ ,  $\ell$ , and for  $n \geq 1$ , find  $\{\bar{\psi}_{c,L,\ell}^n, \bar{\psi}_{w,L,\ell}^n\}$  by solving

$$\phi(x_{L,\ell}) \frac{\bar{\psi}_{c,L,\ell}^n - \bar{\psi}_{c,L,\ell}^{n-1}}{p'_c(s^{n-1})\Delta t^n} - \nabla_{h,L,\ell} \cdot [\bar{\lambda}_w(s^{n-1}) \nabla_{h,L,\ell} \bar{\psi}_w^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m \quad (4.1a)$$

$$-\nabla_{h,L,\ell} \cdot [\bar{\lambda}(s^{n-1}) \nabla_{h,L,\ell} \bar{\psi}_w^n + \bar{\lambda}_o(s^{n-1}) \nabla_{h,L,\ell} \bar{\psi}_c^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m, \quad (4.1b)$$

$$\bar{\psi}_{c,L,\ell}^n = \Psi_{c,L}^{n-1} \quad \text{and} \quad \bar{\psi}_{w,L,\ell}^n = \Psi_{w,L}^{n-1} \quad \text{if } y_\ell \in \partial\mathcal{G}_m, \quad (4.1c)$$

and determine  $\{\check{\psi}_{c,L,\ell}^n, \check{\psi}_{w,L,\ell}^n\}$  and  $\{\hat{\psi}_{c,L,\ell}^n, \hat{\psi}_{w,L,\ell}^n\}$  by solving

$$\phi(x_{L,\ell}) \frac{\check{\psi}_{c,L,\ell}^n}{p'_c(s^{n-1})\Delta t^n} - \nabla_{h,L,\ell} \cdot [\bar{\lambda}_w(s^{n-1}) \nabla_{h,L,\ell} \check{\psi}_w^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m, \quad (4.2a)$$

$$-\nabla_{h,L,\ell} \cdot [\bar{\lambda}(s^{n-1}) \nabla_{h,L,\ell} \check{\psi}_w^n + \bar{\lambda}_o(s^{n-1}) \nabla_{h,L,\ell} \check{\psi}_c^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m, \quad (4.2b)$$

$$\check{\psi}_{c,L,\ell}^n = 1 \quad \text{and} \quad \check{\psi}_{w,L,\ell}^n = 0 \quad \text{if } y_\ell \in \partial\mathcal{G}_m, \quad (4.2c)$$

and

$$\phi(x_{L,\ell}) \frac{\hat{\psi}_{c,L,\ell}^n}{p'_c(s^{n-1})\Delta t^n} - \nabla_{h,L,\ell} \cdot [\bar{\lambda}_w(s^{n-1}) \nabla_{h,L,\ell} \hat{\psi}_w^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m,$$

$$-\nabla_{h,L,\ell} \cdot [\bar{\lambda}(s^{n-1}) \nabla_{h,L,\ell} \hat{\psi}_w^n + \bar{\lambda}_o(s^{n-1}) \nabla_{h,L,\ell} \hat{\psi}_c^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m,$$

$$\hat{\psi}_{c,L,\ell}^n = 0 \quad \text{and} \quad \hat{\psi}_{w,L,\ell}^n = 1 \quad \text{if } y_\ell \in \partial\mathcal{G}_m, \quad (4.3a)$$

where

$$\nabla_{h,L,\ell} \cdot [\bar{\lambda}_\alpha(s^{n-1}) \nabla_{h,L,\ell} \psi^n] = \sum_{j=1}^3 \frac{1}{h_j^2} \left\{ \bar{\lambda}_\alpha \left( \frac{s_{L,\ell+e_j}^{n-1} + s_{L,\ell}^{n-1}}{2} \right) (\psi_{L,\ell+e_j}^n - \psi_{L,\ell}^n) - \bar{\lambda}_\alpha \left( \frac{s_{L,\ell}^{n-1} + s_{L,\ell-e_j}^{n-1}}{2} \right) (\psi_{L,\ell}^n - \psi_{L,\ell-e_j}^n) \right\}.$$

These equations are linear, since the mobilities and  $p'_c$  are evaluated at the previous time level. The matrix potentials  $\psi_{c,L,\ell}^n$  and  $\psi_{w,L,\ell}^n$  are defined below in (4.7) and (4.8); they satisfy the expected equations, namely (4.10). Equations (4.1) define a particular solution to the linear equations, while (4) and (4.3) give solutions to the homogeneous problems which describe unit changes in the boundary conditions.

### iii) The Iterative Procedure.

a) The matrix source term. For each  $L$  and  $n \geq 1$ , compute

$$\begin{aligned} \tilde{\psi}_{c,L,\ell}^{n,i} &= \bar{\psi}_{c,L,\ell}^n + (\Psi_{c,L}^{n,i} - \Psi_{c,L}^{n-1}) \check{\psi}_{c,L,\ell}^n + (\Psi_{w,L}^{n,i} - \Psi_{w,L}^{n-1}) \hat{\psi}_{c,L,\ell}^n, \\ \tilde{s}_{L,\ell}^{n,i} &= p_c^{-1} (\tilde{\psi}_{c,L,\ell}^{n,i} + (\rho_o - \rho_w)gz(x_{L,\ell})), \\ q_{m,w,L}^{n,i} &= -\frac{1}{|\Omega_x|} \sum_{\ell} \phi(x_{L,\ell}) \frac{\tilde{s}_{L,\ell}^{n,i} - s_{L,\ell}^{n-1}}{\Delta t^n} V_{\ell}, \end{aligned} \quad (4.4)$$

where  $V_{\ell}$  is the volume element associated with the grid point  $\ell$ . The quantity  $q_{m,w,L}^n$  is given implicitly in terms of the fracture potentials at the  $n$ th time level; however, in view of (4.7) and (4.9) below, (4.4) is clearly a discretization of (2.4).

b) Fracture System. For each  $L$  and  $n \geq 1$ , solve the nonlinear system of equations (3.5) using the iterative method described in §3 for  $S_L^n$ ,  $Q_{\alpha,L}^n$ ,  $\Psi_{\alpha,L}^n$ ,  $\lambda_{\alpha}^n$ ,  $\alpha = w, o$  by computing  $q_{m,w,L}^{n,i}$  employing iii) above (for simplicity, we denote the set of four Lagrange multipliers associated to each element by  $\lambda_{\alpha}$ ). The no-flow boundary conditions of (3.2b) and (3.2d) are imposed by considering virtual elements outside the computational region such that

$$\lambda_{\alpha}^{n,i} = \lambda_{\alpha,L \mp e_j}^{n,i}, \quad \alpha = w, o, \quad (4.5)$$

$$Q_{\alpha,L \pm e_j}^n = 0, \quad \alpha = w, o, \quad (4.6)$$

if  $x_{L \pm e_j}$  is outside the reservoir.

Since physically meaningful capillary pressures are nonnegative, the capillary functions should be extended vertically downward at  $S_{\max}$  or  $s_{\max}$ .

iv) Matrix update. For each  $L, \ell$ , and  $n \geq 1$ , let

$$\psi_{c,L,\ell}^n = \bar{\psi}_{c,L,\ell}^n + (\Psi_{c,L}^n - \Psi_{c,L}^{n-1}) \check{\psi}_{c,L,\ell}^n + (\Psi_{w,L}^n - \Psi_{w,L}^{n-1}) \hat{\psi}_{c,L,\ell}^n, \quad (4.7)$$

$$\psi_{w,L,\ell}^n = \bar{\psi}_{w,L,\ell}^n + (\Psi_{c,L}^n - \Psi_{c,L}^{n-1}) \check{\psi}_{w,L,\ell}^n + (\Psi_{w,L}^n - \Psi_{w,L}^{n-1}) \hat{\psi}_{w,L,\ell}^n, \quad (4.8)$$

$$s_{L,\ell}^n = p_c^{-1} (\psi_{c,L,\ell}^n + (\rho_o - \rho_w)gz(x_{L,\ell})). \quad (4.9)$$

This completes the time step.

The above algorithm can be implemented sequentially. The following discrete matrix problem has been solved:

$$\phi(x_{L,\ell}) \frac{\psi_{c,L,\ell}^n - \psi_{c,L,\ell}^{n-1}}{p'_c(s^{n-1})\Delta t^n} - \nabla_{h,L,\ell} \cdot [\tilde{\lambda}_w(s^{n-1})\nabla_{h,L,\ell} \psi_w^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m, \quad (4.10a)$$

$$-\nabla_{h,L,\ell} \cdot [\tilde{\lambda}(s^{n-1})\nabla_{h,L,\ell} \psi_w^n + \tilde{\lambda}_o(s^{n-1})\nabla_{h,L,\ell} \psi_c^n] = 0 \quad \text{if } y_\ell \in \mathcal{I}_m, \quad (4.10b)$$

$$\psi_{c,L,\ell}^n = \Psi_{c,L}^n \quad \text{and} \quad \psi_{w,L,\ell}^n = \Psi_{w,L}^n \quad \text{if } y_\ell \in \partial\mathcal{G}_m. \quad (4.10c)$$

Assuming that the wetting fluid is the denser, it should be noted that the block associated with the fracture point  $x_L$  is interpreted to lie below  $x_L$  for imbibition, the case we have treated. For drainage it should be placed above  $x_L$ ; otherwise, fluid is trapped by the numerical simulation as  $P_c$  tends to zero.

The numerical convergence of the iterative method just described is measured in terms of a relative error defined in terms of the  $\ell_2$  norm of variables describing the flow as the number of iterations is successively doubled.

## 5 PARALLEL IMPLEMENTATIONS

We developed a serial code based on the algorithm described in the previous section and validated it against another code developed independently (which uses finite differences) in [16]. In order to minimize execution time we use two techniques to reduce the number of iterations required for convergence. Different procedures are required depending on the time step number. After two time steps have been solved, a quadratic extrapolation in time [14], [13] reduces drastically the number of iterations required for convergence. A different procedure has to be adopted for the two initial time steps. We used a variant of the above method. Instead of extrapolating in time, we consider a spatial hierarchical extrapolation. We solve the problem in a family of nested grids, interpolate the solution according to the finite element method in use on each grid, and then we use a quadratic extrapolation as a function of the grid size. Figure 1 illustrates this procedure. Using the hierarchical extrapolation we typically reduce by a factor of two the execution time for the initial time steps.

We consider speedup studies through simulations of waterflooding calculations in a "five-spot" geometry with gravity effects neglected.

For computational simplicity, the fracture calculations are two-dimensional over  $\Omega$ , though the matrix calculations would remain three-dimensional over each  $\Omega_x$  if gravity were not ignored. Initially, the reservoir contains 75% oil and 25% water. Water is injected uniformly into the reservoir along one corner at a rate of one pore-volume every five years.

The following data are held fixed for the computational results exhibited below:

Fluid properties

Viscosity	$\mu_w = .5 \text{ cP}$	$\mu_o = 2 \text{ cP}$
Density	$\rho_w = 1 \text{ g/cm}^3$	$\rho_o = .7 \text{ g/cm}^3$
Absolute Permeabilities	$K = 1 \text{ darcy}$	$k = 0.05 \text{ darcy}$
Porosities	$\Phi = .01$	$\phi = .1$
Residual Saturations (matrix)	$s_{ro} = .15$	$s_{rw} = .2$
Residual Saturations (fractures)	$S_{ro} = 0$	$S_{rw} = 0$

The capillary pressure functions were assumed in the form

$$\begin{aligned}
 P_c(S) &= (1 - S)\{\gamma(S^{-1} - 1) + \Theta\}, \\
 p_c(s) &= \alpha((s - s_{rw})^{-2} - \beta(1 - s)^{-2}), \\
 s_0 &= 1 - s_{ro}, \quad \beta = s_{ro}^2(s_0 - s_{rw})^{-2}, \\
 \gamma &= 2.0 \times 10^4 \text{ dynes/cm}^2, \quad \Theta = 100 \text{ dynes/cm}^2, \\
 \alpha &= 3.0 \times 10^3 \text{ dynes/cm}^2.
 \end{aligned}$$

The relative permeability functions in the fractures were chosen to be linear, with the residual saturations taken to be zero:

$$K_{ro}(S) = 1 - S, \quad K_{rw}(S) = S.$$

In the matrix blocks the relative permeabilities functions were taken to be

$$\begin{aligned}
 k_{ro}(s) &= \{1 - (1 - s_{ro})^{-1}s\}^2, \\
 k_{rw}(s) &= (1 - s_{rw})^{-2}(s - s_{rw})^2.
 \end{aligned}$$

### 5.1 MIMD Implementation

The MIMD version of the serial code described in the previous sections is implemented through CMMD, a library of the CM-5 and uses SPARC microprocessors. See [27] for additional information about a CM-5 system. A hostless programming model is used, in which each node receives the same copy of a code. The computer code is written in the C language and the driver of the program is written in terms of function pointers. This allows us to assign different functions to distinct subdomains, such that subdomain dependent procedures (like injection of fluid in specified positions and imposition of boundary conditions) can be handled.

The computational domain is decomposed into rectangular regions. Each of the subdomains (which in general will contain several elements of the discretized fracture system of equations) is assigned to a different processor. Each processor allocates memory for the elements contained in its subdomain and for a buffer zone consisting of one layer of elements outside the subdomain. The elements contained in each rectangular region are processed sequentially, using the algorithm described in §3 with a modification which allows exchange of information between nearest neighbor subdomains. Once one step of the iterative procedure is performed on each element within a rectangular



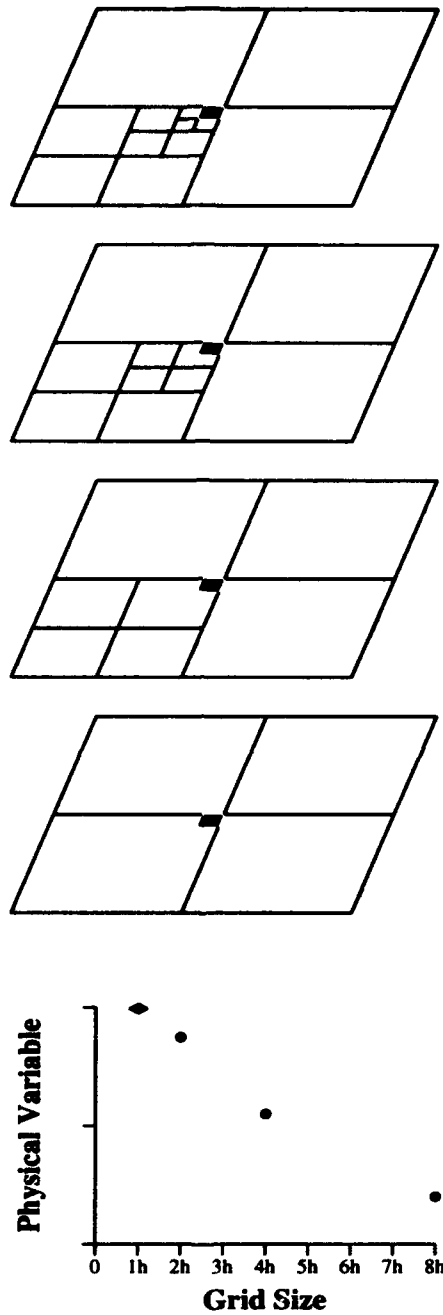
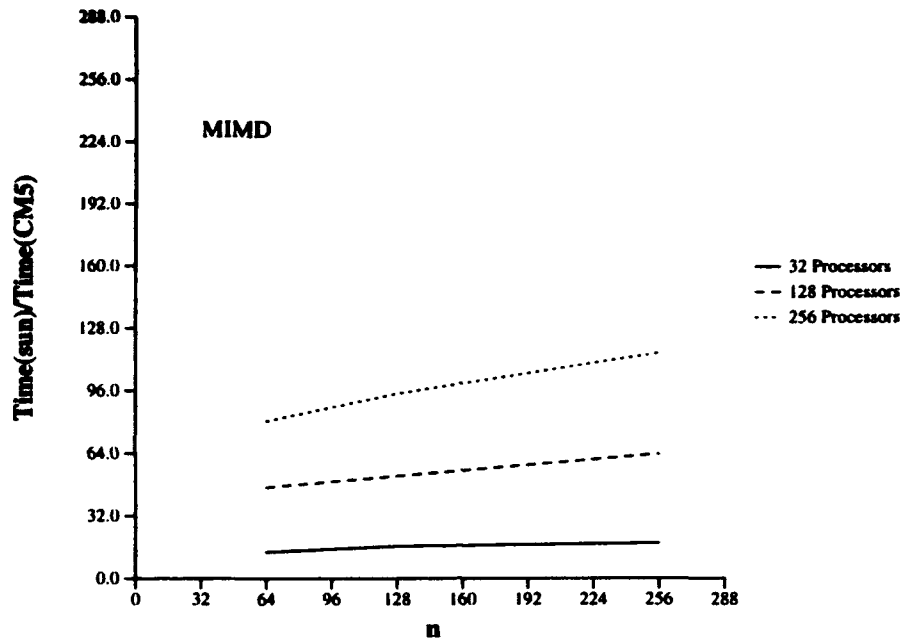


Figure 1: The hierarchical extrapolation. A guess for the iterative method is computed using an extrapolation of solutions of a given problem in coarser grids as a function of the mesh size. Given a problem on a grid with mesh size  $h$  the guess is computed using approximate solutions for the the same problem on grids with sizes  $2h$ ,  $4h$ , and  $8h$ .



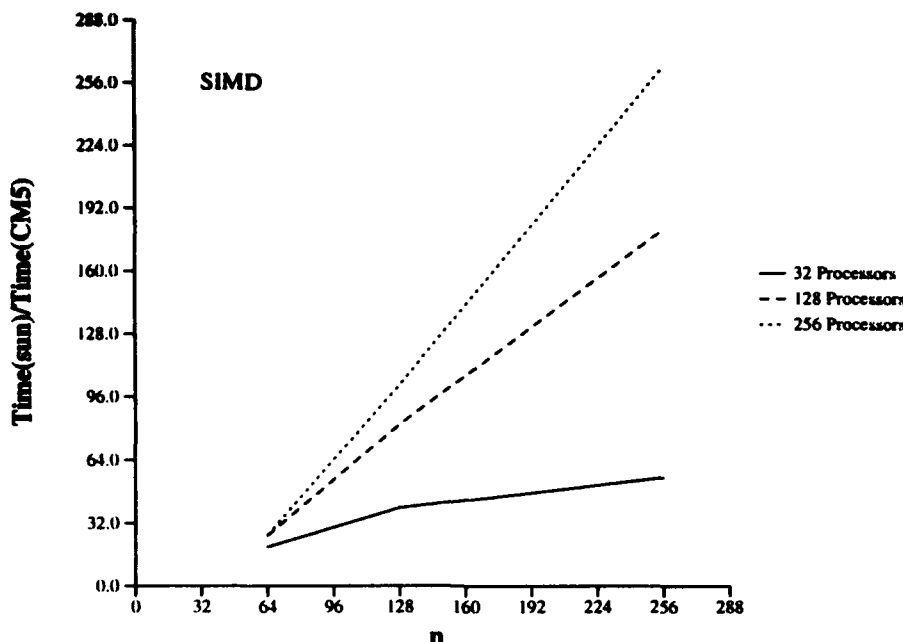
**Figure 2:** The ratio of the time spent by a SUN SPARC station to finish a simulation divided by the time spent by a partition of the CM-5 (running the MIMD version of our code) with variable number of processors is plotted against  $n$ , the number of elements in one direction of the grid. The physical size of the reservoir is increased with  $n$ , keeping the mesh size fixed. For the largest problem considered the speedup obtained is about half of the number of processors used.

region then, through a sequence of grid shifts (right, left, up and down), data on the boundary of subdomains is sent (received) to (from) neighboring subdomains. The boundary conditions (4.5) and (4.6) are also set at this stage of the computation.

We addressed the problem of the speedup obtained with the CM-5 in two studies. First, we compared execution times for simulations performed in the CM-5 with the same simulations performed in a SUN SPARC station. In Figure 2 we plot the ratio of the time spent by a SUN SPARC station to the time spent by different partitions of the CM-5 as a function of the problem size (represented in the plot by the number  $n$  of points in one direction of the grid), with mesh size kept fixed. Note in Figure 2 that for the largest grid considered ( $256 \times 256$ ) the speedup obtained is about half of the number of processors used. Next, we considered the speedup curve for simulations with large grids. We considered the ratio of the time spent by partitions of the CM-5 with 256 and 512 processors (to perform a family of simulations with increasing physical size) to the time spent by 128 processors as a function of the number of processors used. The result of this study is reported in Figure 4. Note in Figure 4 that, as the problem size is increased, the closer the speedup curve gets to the perfect (linear) speedup.

## 5.2 SIMD Implementation

SIMD control in a CM-5 system is achieved through the notion of virtual processors and implemented with data-parallel programming languages. Our code was developed



**Figure 3:** The ratio of the time spent by a SUN SPARC station to finish a simulation divided by the time spent by a partition of the CM-5 (running the SIMD version of our code) with variable number of processors is plotted against  $n$ , the number of elements in one direction of the grid. For problems with small grids increasing the number of processors has little effect on execution time. For the largest problem considered the speedup obtained is more than the number of processors used.

using the language  $C^*$ . Vector processors were used to run the SIMD version of our code.

The  $C^*$  program is quite similar to serial code. We used grid communication within  $C^*$  to perform the necessary exchange of information and the "where" statement to set boundary conditions.

Again, we considered the problem of the speedup obtained with the CM-5 in two studies. An study analogous to the one described in Figure 2 for the MIMD version of our code is the content of Figure 3. Note in Figure 3 that as  $n$  is increased the performance of the CM-5 increases. The speedup obtained for the largest problem size considered ( $256 \times 256$ ) in this study is more than the number of vector processors used. Next, we considered the speedup curve for simulations with large grids. This study appears in Figure 4. As explained above, the ratio of the time spent by partitions of the CM-5 with 256 and 512 processors to the time spent by 128 processors is plotted against the number of processors used. As we noted above for the MIMD version of our code, as the problem size is increased the closer the speedup curve gets to the perfect (linear) speedup.

Figure 4 also allows us to compare the speedup obtained with the two parallel implementations reported here. Although the execution times of the SIMD version are about half of the MIMD version (due to the use of vector processing units) Figure 4 shows a better scalability of the MIMD version.

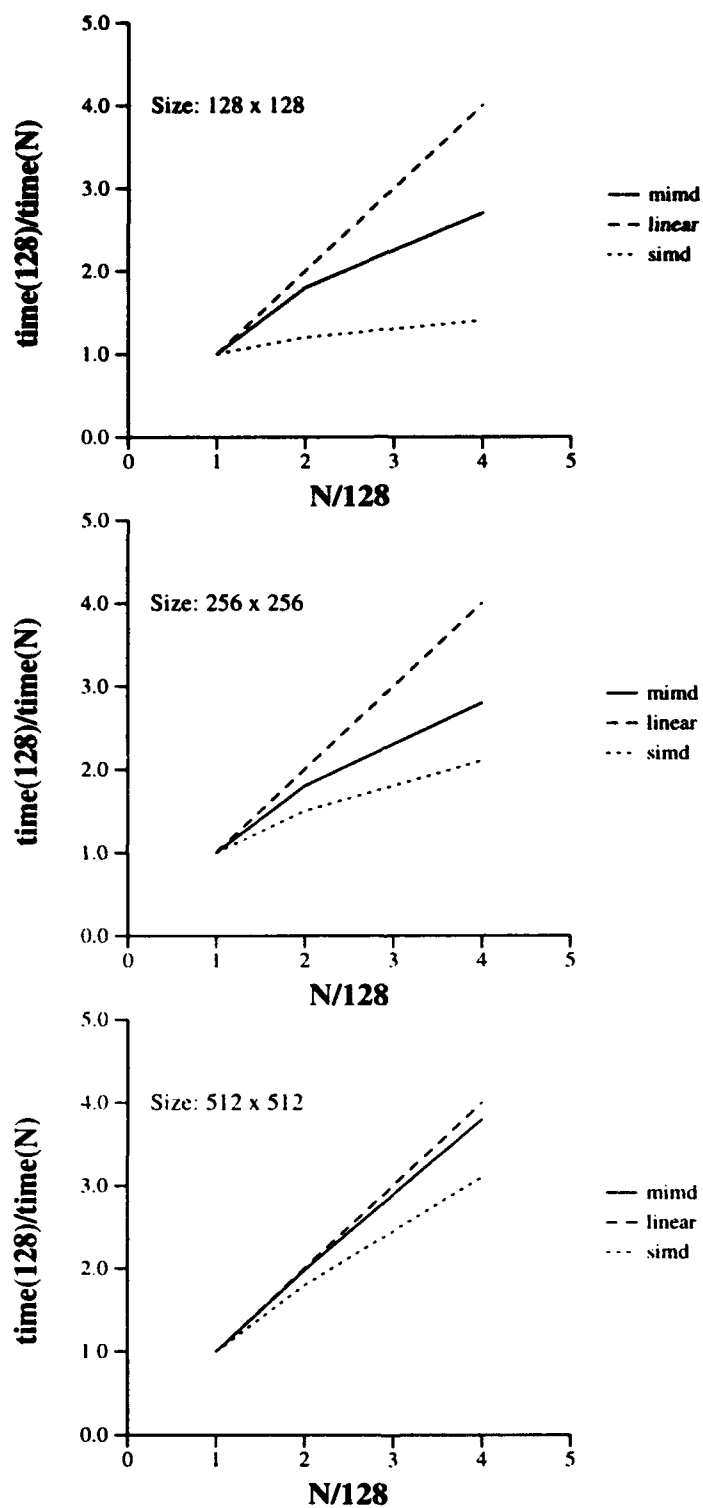


Figure 4: Speedup curves obtained with the MIMD and SIMD versions of the new numerical scheme. As the problem size is increased both versions display a better performance.

## 6 CONCLUSIONS AND OPEN PROBLEMS

We proposed a new numerical method to solve the system of equations governing immiscible flow in naturally fractured reservoirs in massively parallel computers. The numerical method uses spatial decomposition in the context of an iterative procedure to solve the global problem associated with the part of the system governing fracture flow. By decomposing the domain into the elements of the mixed finite element method used in the discretization of fracture equations the numerical scheme can be easily translated into a computer code written in data-parallel languages.

The new numerical procedure was implemented first in a serial machine, and numerical simulations were performed to validate the code. Then, the serial code was restructured and implemented in a CM-5 system, both in MIMD and SIMD modes. We found a good scalability of the two versions of the parallel code for problems with grid sizes relevant in applications.

We established the convergence of the new numerical procedure for a simplified version of the model discussed here, and the proof of convergence for the full system constitutes an interesting open problem. From the numerical point of view, the new numerical method will allow us to study multi-length scale, stochastic, double porosity models, due to the high resolution provided by the CM-5 through the use of fine computational grids. Obviously, full three-dimensional fluid flow simulations remain as one of the most interesting challenges of our research area; such simulations are currently being pursued by the authors.

## REFERENCES

- [1] Arnold D. N., Brezzi F. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO*, 19:7–32, 1985.
- [2] Brezzi F., Douglas J. Jr., Durán R., Fortin M. Mixed finite elements for second order elliptic problems in three variables. *Numer. Math.*, 51:237–250, 1987.
- [3] Brezzi F., Douglas J. Jr., Fortin M., Marini L. D. Efficient rectangular mixed finite elements in two and three space variables. *R.A.I.R.O. Modélisation Mathématique et Analyse Numérique*, 21:581–604, 1987.
- [4] Brezzi F., Douglas J. Jr., Marini L. D. Two families of mixed finite elements for second order elliptic problems. *Numerische Mathematik*, 47:217–235, 1985.
- [5] Brezzi F., Douglas J. Jr., Marini L. D. Variable degree mixed methods for second order elliptic problems. *Matemática Aplicada e Computacional*, 4:19–34, 1985.
- [6] Chen Z., Douglas J. Jr. Prismatic mixed finite elements for second order elliptic problems. *Calcolo*, 26:135–148, 1989.
- [7] Cowsar L. C., Wheeler M. F. Parallel domain decomposition method for mixed finite elements for elliptic partial differential equations. In *Proceedings of the Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations*, Philadelphia, 1991. SIAM. R. Glowinski, Y. Kuznetsov, G. Meurant, J. Périaux, and O. B. Widlund, eds.

- [8] Després B. *Domain decomposition method and the Helmholtz problem*, pages 44–52. SIAM, Philadelphia, 1991. G. Cohen, L. Halpern, and P. Joly (eds.).
- [9] Després B. *Méthodes de décomposition de domaines pour les problèmes de propagation d'ondes en régime harmonique*. PhD thesis, Université Paris IX Dauphine, UER Mathématiques de la Décision, 1991.
- [10] Després B., Joly P., Roberts J. E. *A domain decomposition method for the harmonic Maxwell equations*, pages 475–484. Elsevier Science Publishers B. V. (North-Holland), Amsterdam, 1992. R. Beauwens and P. de Groen (eds.).
- [11] Douglas J., Jr., Arbogast T. Dual porosity models for flow in naturally fractured reservoirs. In *Dynamics of Fluids in Hierarchical Porous Formations*, pages 177–221. Academic Press, London, 1990. J. H. Cushman, ed.
- [12] Douglas J. Jr., Arbogast T., Paes Leme P. J., Hensley J. L., Nunes N. P. Immiscible displacement in vertically fractured reservoirs. *Transport in Porous Media*. To appear, 1993.
- [13] Douglas J. Jr., Dupont T., Ewing R. E. Incomplete iteration for time-stepping a nonlinear parabolic Galerkin method. *SIAM J. Numer. Anal.*, 16:503–522, 1979.
- [14] Douglas J. Jr., Dupont T., Percell P. A time-stepping method for Galerkin approximations for nonlinear parabolic equations. In *Numerical Analysis, Dundee 1977*, volume 630 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1978.
- [15] Douglas J. Jr., Hensley J. H., Paes Leme P. J. A study of the effect of inhomogeneities on immiscible flow in naturally fractured reservoirs. In *Porous Media*. Birkhäuser, Basel, 1993.
- [16] Douglas J., Jr., Hensley J. L., Arbogast T. A dual-porosity model for waterflooding in naturally fractured reservoirs. *Computer Methods in Applied Mechanics and Engineering*, 87:157–174, 1991.
- [17] Douglas J. Jr., Paes Leme P. J., Roberts J. E., Wang J. A parallel iterative procedure applicable to the approximate solution of second order partial differential equations by mixed finite element methods. *Numerische Mathematik*. To appear, 1993.
- [18] Ewing R. E., Wang J. Analysis of multilevel decomposition iterative methods for mixed finite element methods. *R.A.I.R.O. Modélisation Mathématique et Analyse Numérique*. Submitted.
- [19] Ewing R. E., Wang J. Analysis of the Schwarz algorithm for mixed finite element methods. *R.A.I.R.O. Modélisation Mathématique et Analyse Numérique*, 26:739–756, 1992.
- [20] Fraeijs de Veubeke B. X. Stress function approach. International Congress on the Finite Element Method in Structural Mechanics, Bournemouth, 1975.
- [21] Fraeijs de Veubeke B. X. Displacement and equilibrium models in the finite element method. In *Stress Analysis*. John Wiley, New York, 1965. O. C. Zienkiewicz and G. Holister (eds.).

- [22] Glimm J., Lindquist B., Pereira F., Peierls R. The fractal hypothesis and anomalous diffusion. *Matemática Aplicada e Computacional*, 11:189–207, 1992.
- [23] Glimm, J., Lindquist B., Pereira F., Zhang Q. A theory of macrodispersion for the scale up problem. *Advances in Water Resources*. To appear.
- [24] Glowinski R., Kinton W., Wheeler M. F. Acceleration of domain decomposition algorithms for mixed finite elements by multi-level methods. In *Third International Symposium on Domain Decomposition Methods for Partial Differential Equations*, pages 263–290. SIAM, Philadelphia, 1990. R. Glowinski (ed.).
- [25] Glowinski R., Wheeler M. F. *Domain decomposition and mixed finite element methods for elliptic problems*, pages 144–172. SIAM, Philadelphia, 1988. R. Glowinski, G. Golub, G. Meurant, and J. Periaux (eds.).
- [26] Nedelec J. C. Mixed finite elements in  $R^3$ . *Numer. Math.*, 35:315–341, 1980.
- [27] Palmer J., Steele G. L. Jr. Connection Machine Model CM-5 system overview. In *Proceedings of the Fourth Symposium on the Frontiers of Massively Parallel Computation*, pages 474–483. IEEE Computer Society Press, Los Alamitos, 1992.
- [28] Raviart P. A., Thomas J. M. A mixed finite element method for second order elliptic problems. In *Mathematical Aspects of the Finite Element Method*, volume 606 of *Lecture Notes in Mathematics*, pages 292–315. Springer-Verlag, Berlin and New York, 1977. I. Galligani and E. Magenes, eds.
- [29] Thomas J. M. *Sur l'analyse numérique des methodes d'éléments finis hybrides et mixtes*. PhD thesis, Université Pierre-et-Marie Curie, Paris, 1977.

# NUMERICAL SOLUTION OF RICHARDS' EQUATION

A. J. SILVA NETO<sup>1</sup> and R. E. WHITE

Department of Mathematics  
North Carolina State University  
Raleigh, NC 27695-8205, USA

## SUMMARY

The model for fluid flow in partially saturated porous media involves a nonlinear parabolic partial differential equation for the hydraulic head pressure, i.e. Richards' equation. We focus on problems with large derivatives of the moisture content and hydraulic conductivity functions. Here similarities with the Stefan problem yield the successful application of the nonlinear SOR method to Richards' equation. Implementation issues for vector and multiprocessor computers are also discussed.

## INTRODUCTION

Richards(1931) developed a nonlinear parabolic equation that models fluid flow in porous media. Richards' equation relies on empirical nonlinear functions such as the moisture content,  $\Theta(h)$ , and the hydraulic conductivity,  $K(h)$ :

$$\Theta(h)_t - \nabla \cdot K(h) \nabla h - K(h)_z = 0 \quad (1a)$$

where  $h$  is the hydraulic pressure head, and  $z$  is the vertical direction.

The boundary and initial conditions usually have the form

$$[K(h) \nabla (h + z)] \cdot n = \text{given for the boundary} \quad (1b)$$

$$h = \text{given for } t = 0 \quad (1c)$$

where  $n$  is a unit outward normal to the space domain.

Richards' equation coupled to a parabolic system for the contaminants becomes the basic model for the analysis of transport of contaminants in groundwater (Freeze and Cherry, 1979; Feng, 1993).

Traditional methods (Paniconi et al, 1991) of solution of problem (1) such as Newton, Picard and Lees implicit factored schemes, seem to work well for problems without large derivatives in  $\Theta(h)$  and  $K(h)$ . But, some oscillations can develop for the large derivative cases.

Here we will approximate the empirical functions  $\Theta(h)$  and  $K(h)$  by piecewise linear functions. This approach is justified by the often significant uncertainties related to the computation of  $\Theta(h)$  and  $K(h)$ . The computation times are reduced by avoiding the calculation of exponential terms (van Genuchten and Nielsen, 1985), and one can still use the monotonicity of  $\Theta(h)$  and  $K(h)$  to analyse convergence and obtain bounds for the solutions (i.e. comparison results).

In this work we show the similarities between the enthalpy formulation of the Stefan problem and the analogous moisture

formulation of the Richards' equation. Then we focus on the solution of problem (1) using a nonlinear SOR set-valued algorithm (Cryer, 1971). We will also discuss high performance computing issues.

## MOISTURE FORMULATION OF RICHARDS' EQUATION

The empirical functions of  $\Theta(h)$  and  $K(h)$  are given by van Genuchten and Nielsen (1985) in the form

$$\Theta(h) = \begin{cases} \Theta_r + (\Theta_s - \Theta_r)(1 + |\alpha h|^n)^{-m}, & h < h_0 \\ \Theta_r + (\Theta_s - \Theta_r)(1 + |\alpha h_0|^n)^{-m} + s_r(h - h_0), & h \geq h_0 \end{cases} \quad (2a)$$

$$+ s_r(h - h_0), \quad h \geq h_0 \quad (2b)$$

$$K(h) = \begin{cases} K_s S(h)^{\frac{1}{2}} [1 - (1 - S(h)^{\frac{1}{2}})^m]^2, & h < h_0 \\ K_s, & h \geq h_0 \end{cases} \quad (3a)$$

$$, \quad h \geq h_0 \quad (3b)$$

$$S(h) = \begin{cases} \frac{\Theta - \Theta_r}{\Theta_s - \Theta_r} = (1 + |\alpha h|^n)^{-m}, & h < h_0 \\ 1.0, & h \geq h_0 \end{cases} \quad (4a)$$

$$, \quad h \geq h_0 \quad (4b)$$

where the parameters  $\Theta_r$ ,  $\Theta_s$ ,  $n$ ,  $m = 1 - \frac{1}{n}$ ,  $\alpha$ ,  $K_s$  and  $s_r$  are function of the particular porous media, and  $h_0$  is chosen so that  $\frac{d\Theta}{dh}$  is continuous at  $h_0$ . Often  $\Theta(h)$  have large derivatives, and  $\frac{d\Theta}{dh}$  appears to be discontinuous in numerical calculations. Also large uncertainties are associated with functions in (2) to (4) because of the unknown nature of the subsurface soil.

When the hydraulic conductivity,  $K$ , is a function of  $h$  only, a Kirchhoff transformation

$$v = F(h) \equiv \int_0^h K(h) dh \quad (5)$$

is used and Richards' equation becomes

$$\Theta(F^{-1}(v))_t - \nabla^2 v - K(F^{-1}(v))_z = 0 \quad (6)$$

When  $\Theta$  is viewed as the primary unknown, Eq.(6) is called the moisture formulation of Richards' equation.

<sup>1</sup>Permanent address: Promon Engenharia, Av. Pres. Juscelino Kubitschek 1830, São Paulo, SP, 04543-900, Brasil



This is analogous to the enthalpy formulation of the Stefan problem where  $\Theta$  is replaced by the enthalpy and  $h$  by the temperature. In the Stefan problem the  $K_s$  term would be equivalent to a heat source at the phase change interface.

For the remainder of this paper we assume two space dimensions, with  $y$  replacing  $z$  on the vertical direction.

### THE FINITE DIFFERENCE METHOD

As in the Stefan problem for long time durations, the implicit time discretisation is used in Eq.(6),

$$\frac{\Theta^{m+1} - \Theta^m}{\Delta t} - \nabla^2 v^{m+1} - K(v^{m+1})_v = 0 \quad (7)$$

The FDM for the interior grid points is

$$d_{ij} - c v_{ij} = \frac{\Theta(F^{-1}(v_{ij}))}{\Delta t} + \frac{K(F^{-1}(v_{ij}))}{\Delta y} \quad (8)$$

where

$$d_{ij} \equiv \frac{K(F^{-1}(v_{i,j+1}))}{\Delta y} + \frac{1}{\Delta x^2}(v_{i-1,j} + v_{i+1,j}) + \frac{1}{\Delta y^2}(v_{i,j-1} + v_{i,j+1}) + \frac{1}{\Delta t} \bar{\Theta}_{ij} \quad (9)$$

$$c = \frac{2}{\Delta x^2} + \frac{2}{\Delta y^2} \quad (10)$$

$$\bar{\Theta}_{ij} = \Theta_{ij}^m + \Delta t \left( \bar{d}_{ij} - \frac{K(F^{-1}(\bar{v}_{ij}))}{\Delta y} - c \bar{v}_{ij} \right) \quad (11)$$

and  $\bar{v}_{ij} = v_{ij}$  from the previous time step.

Similar equations are written for the cells at the boundaries.

If there are jump discontinuities in either  $\Theta(h)$  or  $K(h)$ , Eq.(8) is a set-valued equation in the form

$$d - c v \in \Gamma(v) \quad (12)$$

Here we are trying to find  $v$  so that  $d - c v$  is an element of the set  $\Gamma(v)$ . As depicted in Fig.1, there are three cases to consider. For each case there is a solution, and it is unique. As in the enthalpy formulation of the Stefan problem, this will accurately track the moisture in regions where large derivatives or jump discontinuities are located.

### NONLINEAR SOR FOR SET-VALUED SYSTEMS

The following algorithm is applied to Eq.(8) or to the set-valued system (12), where  $maxit$  = maximum allowable SOR iterations,  $n_x, n_y$  = number of cells in the  $x$  and  $y$  directions respectively,  $1.0 \leq \bar{\omega} < 2.0$  is the SOR parameter which is to be applied to either the unsaturated or the saturated cells, and  $0.8 \leq \underline{\omega} \leq 0.9$  is used to dampen numerical oscillations.

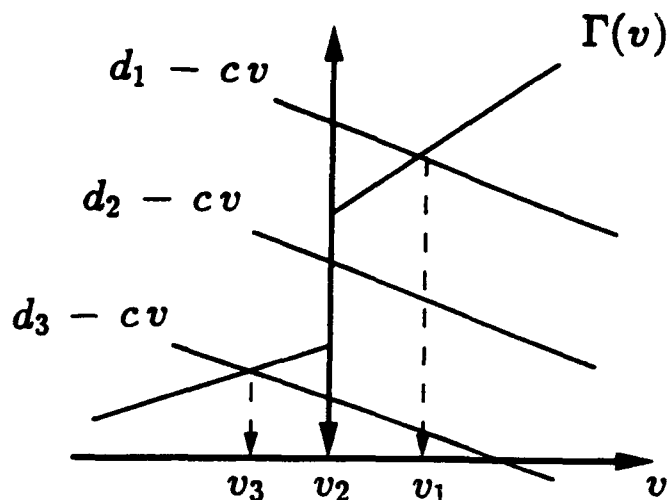


Figure 1. Set-valued Equation.

### Nonlinear SOR Algorithm.

```

for k = 1, maxit
  for i = 1, n_x
    for j = 1, n_y
      compute dij from Eqs.(8) or (12)
      solve dij - c vij ∈ Γ(vij) as in Fig.1
      if v vij > 0, then
        vij = (1 -  $\bar{\omega}$ )vij +  $\bar{\omega}$  v
      else
        vij = (1 -  $\underline{\omega}$ )vij +  $\underline{\omega}$  v
      end if
    end loop j
  end loop i
  check for convergence
end SOR loop k

```

In our calculations the stopping criterion consisted of requiring the difference for calculated  $v$  between successive iterations to be less than  $\epsilon h^2$ , where  $h = \max(\Delta x, \Delta y)$  and  $\epsilon$  ranged from  $10^{-3}$  to  $10^{-5}$ . Convergence was usually obtained in 10 to 30 iterations.

### HIGH PERFORMANCE COMPUTING

We tried vectorization of the CRAY Y-MP computer, vectorization and multiprocessing on a two processor ALLIANT FX-40, and multiprocessing on the Kendall Square Research KSR1 computer.

In general the vectorization did not work well. Here we used Red-Black ordering (White,1987), but the computations in the inner most loops were too complicated to vectorize effectively. This was observed on both the CRAY Y-MP and the ALLIANT FX-40.

In order to use multiprocessor, a domain-decomposition technique (White,1987) is applied, as illustrated in Fig.2. We ordered the even blocks first and then the odd blocks. The nonlinear SOR algorithm is then executed in parallel as follows.

```

for k = 1, maxit
  concurrently do nonlinear SOR over the even blocks
  update
  concurrently do nonlinear SOR over the odd blocks
  update
  check for convergence
end SOR loop k

```

The KSRI computer has three parallel constructs that can be used in FORTRAN code: *tiles*, *parallel sections* and *parallel regions*. *Tiles* are used to partition loops and are very effective for simple cases such as matrix multiplications. *Parallel sections* can be used to concurrently execute different code segments. We used *parallel regions*, in which code segments are duplicated to allow concurrent computations. In our implementation we used *teams of processors* that are assigned in the initial part of the program due to the high overhead.

Table 1 shows a comparison of the CPU time on the KSRI computer of the North Carolina Supercomputing Center for the computation of the first time step of Example 3 presented in the next section. Here  $N = n_x = n_y$  is the number of grid cells in each direction,  $L$  is the number of large blocks ( $L=4$  in Fig.2),  $S_L = \frac{\text{CPU time using one block}}{\text{CPU time using } L \text{ blocks}}$  is the speedup, and  $E_L = \frac{S_L}{L}$  is the efficiency.

Table 1. Speedup for Domain-Decomposition SOR on KSRI

$N$	$L$	time (s)	$S_L$	$E_L$
80	1	8.74	1.00	1.00
80	2	5.57	1.57	0.79
80	4	3.02	2.89	0.72
80	8	2.22	3.94	0.49
160	1	49.49	1.00	1.00
160	2	27.75	1.78	0.89
160	4	14.41	3.43	0.86
160	8	8.73	5.67	0.71

We observed a declining efficiency,  $E_L$ , with an increasing number of large blocks,  $L$ . This is consistent with Amdahl's law (Ortega, 1988). A second important observation is the increase in efficiency with the increase on the number of grid cells,  $N$ . This fact is explained by the relatively smaller ratio of amount of parallel overhead to the amount of concurrent computations.

These results were also observed for other computations and indicates that the KSRI computer can be used efficiently for larger problems.

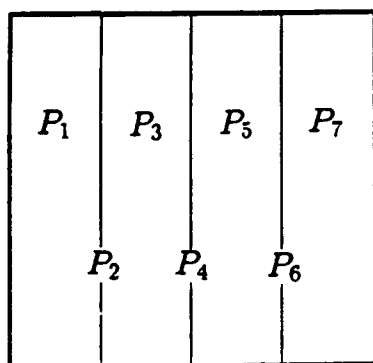


Figure 2. Domain-Decomposition.

To validate the numerical model we present the results for three numerical experiments. For the first two, analytical solutions are available, and for the third one a comparison with experimental data is done.

For all computations presented here we used piecewise linear approximations of the empirical functions (2)-(4) for the moisture content,  $\Theta(h)$ , and hydraulic conductivity,  $K(h)$ , as shown in Fig.3. These approximations allow the test of the method under most strict conditions due to the discontinuity of the derivatives with respect to the pressure head. In fact it will also represent savings on CPU time by avoiding the quite involved calculations required in Eqs.(2) to (4).

The first and second examples are in one space dimension and have the boundary and initial conditions in the form

$$h(y, t) = e^t - 1 \quad \text{at } y = 0, \quad (13a)$$

$$h(y, t) = \alpha (e^{t-1} - 1) \quad \text{at } y = 1, \quad (13b)$$

$$h(y, t) = \alpha (e^{-y} - 1) \quad \text{for } t = 0. \quad (13c)$$

The analytical solutions for particular choices of  $\alpha$  and the parameters in Fig.3, have the form

$$h = \begin{cases} e^{t-y} - 1 & , t > y \\ \alpha (e^{t-y} - 1) & , t \leq y \end{cases} \quad (14a)$$

$$(14b)$$

**Example 1.** This example corresponds to a situation in which there is no water supply from the top of the domain and there is an infiltration from the bottom, as from a groundwater source.

The parameters used in the piecewise linear approximation of  $\Theta(h)$  and  $K(h)$ , as shown in Fig.3, were  $\alpha_1 = 1$ ,  $\theta_1 = 0$ ,  $\alpha_2 = 2$ ,  $\theta_2 = 1$ ,  $c_1 = 0$ ,  $k_1 = 1$ ,  $c_2 = 0$  and  $k_2 = 2$ . It is considered also  $\alpha = 0$  in Eqs.(13).

In the numerical experiments we used 20 and 40 grid cells, and  $\Delta t = \Delta y$ . Convergence was uniformly rapid and the interface was always located to within one grid cell (i.e. in all runs we were able to keep track of the saturated/unsaturated interface very accurately).

Some oscillations were noted near the interface, which is similar to what happens with the enthalpy method for the Stefan problem. In fact the problem at hand can be viewed as a one phase Stefan problem with a source at the interface.

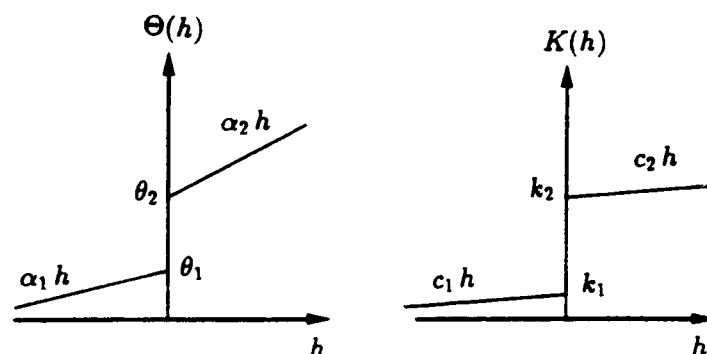


Figure 3. Piecewise Linear Approximation of  $\Theta(h)$  and  $K(h)$ .

**Example 2.** This example corresponds to a situation in which there is a water supply from both the top and the bottom of the domain.

The parameters used in the piecewise linear approximation of  $\Theta(h)$  and  $K(h)$  were  $\alpha_1 = \frac{4}{3}$ ,  $\theta_1 = 0$ ,  $\alpha_2 = 1$ ,  $\theta_2 = 1$ ,  $c_1 = 0$ ,  $k_1 = \frac{4}{3}$ ,  $c_2 = 0$  and  $k_2 = 2$ . It is considered also  $\alpha = \frac{1}{4}$  in Eqs.(13).

Similar performance to Example 1 was observed.

**Example 3.** The repacked Brindabella silty loam has data corresponding to Eqs.(2) to (4) as given by Feng(1993):  $K_s = 0.118$  m/h,  $\Theta_s = 0.110$ ,  $\Theta_s = 0.485$ ,  $\alpha = 2.857$ ,  $n = 1.8$ ,  $s_s = 0$ , and  $h_0 = 0$ .

The flux at the top boundary was taken as 0.0165 m/h, while the other boundary segments had zero flux. The initial condition was -8.0 m, and the region was 0.3 m  $\times$  0.3 m. We attempted to reproduce the data given by White and Broadbridge(1988); therefore, we focused on the portion of  $\Theta(h)$  and  $K(h)$  such that  $\Theta \leq 0.425$ . Here we used a coarse approximation of the empirical functions  $\Theta(h)$  and  $K(h)$  adopting  $\alpha_1 = 0.021$ ,  $\theta_1 = 0.275$ ,  $\alpha_2 = 0.0$ ,  $\theta_2 = 0.425$ ,  $k_2 = 0.0165$ ,  $c_1 = 0.0$ ,  $c_2 = 0.0$ , and  $k_1 = 0.00001$ . The results are given in Fig.4.

The nonlinear SOR algorithm converged with no difficulties until saturation was reached. The differences in the data and the computed values are attributed to the crude approximation of the hydraulic conductivity function.

For more complicated hydraulic conductivity functions, one should not use the Kirchhoff transformation, Eq.(5). There are two reasons for this: First, the solution of the SOR step in Eq.(8) or (12) becomes more complicated. Second, in most porous media the hydraulic conductivity is also space dependent, in which case the Kirchhoff transformation does not simplify the diffusion part of Richards' equation.

## CONCLUSIONS

The nonlinear SOR algorithm was coupled with the moisture formulation to approximate the solution of Richards' equation. Here special attention was given to problems with large derivatives of the moisture content and hydraulic conductivity functions. The nonlinear SOR algorithm converged rapidly for partially saturated regions. Moreover, it did adapt very well to multiprocessing, but not so well to vectorization.

Although the calculations in this paper were for one and two space variables and for homogeneous porous media, there should be little problem in generalizing the method to three space variables and to inhomogeneous media with an assortment of nonlinear terms.

## ACKNOWLEDGEMENTS

The authors wish to acknowledge the North Carolina Supercomputing Center for providing time on CRAY Y-MP and Kendall Square computers. A.J.S.N. acknowledges the financial support given by CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico and Promon Engenharia.

## REFERENCES

Cryer, C.W., "The Solution of a Quadratic Programming Problem Using Systematic Overrelaxation", *SIAM J. Control*, Vol.9, No3, pp.385-392, 1971.

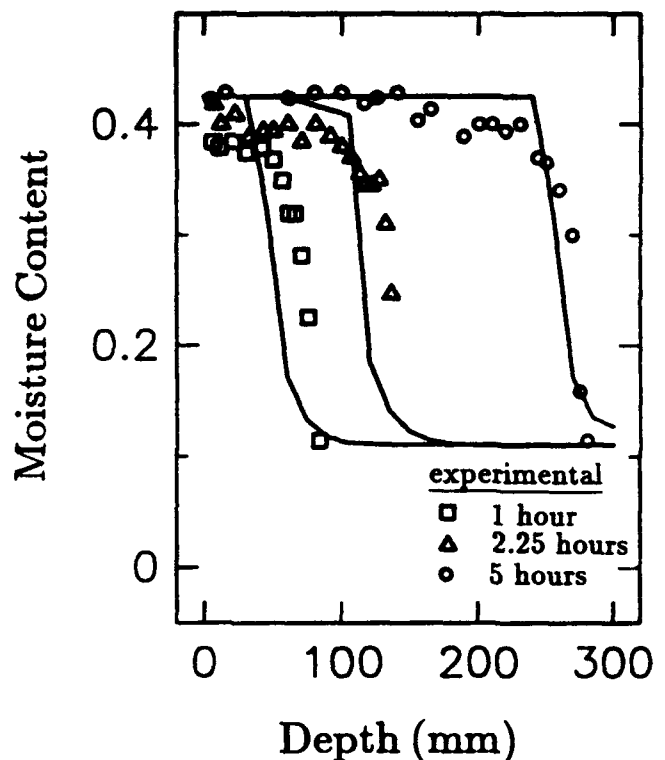


Figure 4. Moisture Content of a Silty Loam.

- Feng, J., *Modeling of Chemical Transport from Agricultural Wast. Lagoons*, Ph.D. Dissertation, North Carolina State University, USA, 1993.
- Freeze, R.A. and Cherry, J.A., *Groundwater*, Prentice Hall Inc, Englewood Cliffs, N.J., 1979.
- Ortega, J.M., *Introduction to Parallel and Vector Solution of Linear Systems*, Plenum Press, New York, 1988.
- Paniconi, C., Aldama, A.A. and Wood, E.F., "Numerical Evaluation of Iterative and Noniterative Methods for the Solution of the Nonlinear Richards Equation", *Water Resources Research*, Vol.27 No6, pp.1147-1163, 1991.
- Richards, L.A., "Capillary Conduction of Liquids through Porous Media", *Physics*, Vol.1, pp.318-333, 1931.
- van Genuchten, M.T. and Nielsen, D.R., "On Describing and Predicting the Hydraulic Properties of Unsaturated Soils", *Ann. Geophys.*, Vol.3, No5, pp.615-628, 1985.
- White, I. and Broadbridge, P., "Constant Rate Rainfall Infiltration: A Versatile Nonlinear Model. 2. Applications of Solutions", *Water Resources Research*, Vol.24, No1, pp.155-162, 1988.
- White, R.E., "Multisplittings and Parallel Iterative Methods", *Computer Methods in Applied Mechanics and Engineering*, Vol.64 pp.567-577, 1987.

# NUMERICAL SOLUTION OF FLUID FLOW IN PARTIALLY SATURATED POROUS MEDIA<sup>1</sup>

by

A. J. Silva Neto<sup>2</sup>

R. E. White<sup>3</sup>

Department of Mathematics  
Box 8205  
North Carolina State University  
Raleigh, NC 27695-8205

## ABSTRACT

This paper describes an SOR algorithm for solving the nonlinear algebraic system which evolves from Richards' equation that models fluid flow in a porous media. The moisture content and hydraulic conductivity functions are approximated by piecewise linear functions obtained from field data. The resulting algebraic system is solved by a variation of the nonlinear SOR algorithm. The advantage of this approach is that it avoids some of the numerical oscillations associated with large derivatives in the data. Numerical calculations are presented and illustrate the following: (i) agreement of the numerical model with observed data, (ii) dependence and comparison results as a function of uncertain data, and (iii) suitability of these algorithms for multiprocessing computations via domain decomposition methods. Extension of these algorithms to heterogeneous porous media fluid flow are discussed.

---

<sup>1</sup>The calculations were done at the North Carolina Supercomputing Center.

<sup>2</sup>Supported by CNPq and Promon Engenharia from Brazil.

<sup>3</sup>Supported by U. S. ARO contract number DAAL03-90-G-0126.

## 1. INTRODUCTION

The study of fluid flow in porous media has several important applications in engineering (Kaviany 1991 and Nield and Bejan 1992). Specific examples are: filters for industrial use, or separators in aerospace fuels (Kaviany 1991); use of geothermal energy (Rae et al. 1983 and Kimura 1989, 1989a); oil recovery (Bear 1972); groundwater (Mark 1992, 1993 and Clothier et al. 1981) and agriculture (Feng 1993).

Industrial chemical or radioactive effluents are sometimes deposited at the surface or in drums that are buried underground. In both normal operations and in accidental conditions, it is required to give an analysis of the transport of the contaminants through the soil (Muralidhar 1990, 1993). The first step in this analysis is the mathematical simulation of fluid flow through the soil.

Richards (1931) developed an equation that is a combination of the continuity equation and Darcy's law (Philip 1969), and it models fluid flow in a porous medium. It is a nonlinear parabolic partial differential equation which contains the empirical functions for *moisture content*  $\theta(h)$  and *hydraulic conductivity*  $K(h)$ .

$$\theta(h)_t - \nabla \cdot K(h) \nabla h - K(h)_z = 0 \quad \text{in } \Omega \times (0, T) \quad (1a)$$

where  $h$  is the hydraulic pressure head,  $z$  is the vertical direction and  $\Omega$  is the space domain. The boundary condition of the third kind has the form

$$|K(h) \nabla (h + z)| \cdot n = \text{given} \quad \text{in } S \times (0, T) \quad (1b)$$

where  $n$  is the unit outward normal to  $\Omega$  and  $S$  is the boundary of  $\Omega$ . The initial condition is

$$h = \text{given} \quad \text{for } t = 0 \quad \text{in } \Omega. \quad (1c)$$

In general the equations (1a-1c) are coupled with a parabolic system of equations for the transport of a number of contaminants through the soil (Freeze and Cherry 1979 and Feng 1993). This is done by using the fluid velocity  $v = K(h)\nabla(h + z)$  which is computed from the above system.

In practice the empirical functions for moisture content and hydraulic conductivity have several troublesome properties. First, they can have large derivatives, and this is often the case for hydraulic conductivity. Second, they are not precisely known. Third, they can have strong space dependence with jump discontinuities resulting from heterogeneous porous media. The objective of this paper is to give an approach to these problems which is based on methods used for the Stefan problem (Silva Neto and White 1993). Particular attention will be given to the first problem where there is no space dependence. In the case of space dependent empirical functions, one can use additional nodes and the continuity condition on the fluid velocity (White et al. 1993) to generalize the methods of this paper.

Traditional methods for the solution of (1a-1c) use an approximation of the empirical data by exponential functions (van Genuchten and Nielsen 1985). Then numerical methods such as Newton, Picard, or Lees implicit factored method can be used for problems without large derivatives (Paniconi et al. 1991). In addition to addressing the above problems, the approach of this paper does not involve expensive function evaluations and does eliminate the numerical oscillations associated with large derivatives of the empirical functions.

In section two we present the general approach to the problem which is adapted from the Stefan problem. Here the empirical functions are approximated by piecewise linear functions which reflect the field data (White and Broadbridge 1988). The partial differential equation is discretized by the finite difference method, and the resulting nonlinear system is solved by a nonlinear SOR algorithm (Cryer 1971) which is described in section three.

Numerical experiments are presented in sections four, five and six, and these experiments were chosen to demonstrate the feasibility of realistic two dimensional simulation of porous flows. Later, we indicate how one can extend these methods to three dimensional and heterogeneous porous flows. We show agreement of the numerical model with the field data from Brindabella silty loam soil. Also, we show how one can develop comparison results which deal with the uncertain empirical data. High performance computing issues are described. Here we demonstrate that the algorithm in section three does not vectorize well, but it does work well for multiprocessors when domain decomposition methods are used. Finally, we state our conclusions and related work.

## **2. DISCRETE VERSION OF RICHARDS' EQUATION**

In this section we state the finite difference discretization of Richards' equation and make some comparisons with the Stefan problem. If in equation (1a) the last term is eliminated, and  $h$  were to represent temperature with  $K$  now denoting the thermal conductivity and  $\theta$  the enthalpy, then this would be the enthalpy formulation of the Stefan problem (White 1985). In the Stefan problem the  $K$  and  $\theta$  have jump discontinuities at the phase change temperature. SOR methods can be effectively used provided the overrelaxation is not applied during a cell's phase change.

In Richards' equation we will approximate  $K$  and  $\theta$  by piecewise linear functions which could be viewed as a number of "linear phases" associated with the nonlinear flow. As in the Stefan problem we will apply the SOR method provided the cell is not changing "linear phase." Table 1 gives some data for Brindabella loam which was extrapolated from the graphs in White and Broadbridge (1988). Note, both  $\theta(h)$  and  $K(h)$  are monotone, and  $K(h)$  has large derivatives.

**Table 1: Brindabella Data**

<b>h</b> <b>[m]</b>	<b><math>\theta(h)</math></b> <b>[fraction]</b>	<b>K(h)</b> <b>[m/hr]</b>
-8.0	0.11	0.0
-1.2	0.27	0.000 118
-0.8	0.30	0.000 327
-0.7	0.31	0.000 457
-0.6	0.32	0.000 664
-0.5	0.33	0.001 138
-0.4	0.34	0.001 693
-0.3	0.36	0.002 326
-0.2	0.38	0.004 568
-0.1	0.42	0.011 117
-0.0	0.485	0.118 000

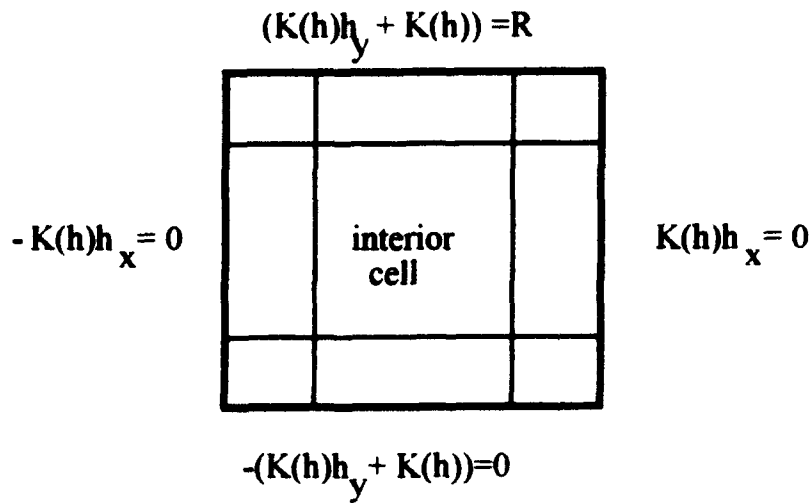
In Silva Neto and White (1993) two "linear phases" were used and were coupled with the Kirchhoff change of dependent variable. Although this was a crude approximation of the data, it did track the wet-dry interface where a rapid transition from unsaturated to saturated regions occurs. In cases where either the transition is not rapid or there is space dependence of the data, the Kirchhoff transformation is not applicable. In the following we use an implicit time discretization of Richards' equation.

$$\frac{\theta(h) - \theta(\bar{h})}{\Delta t} - (K(h)h_x)_x - (K(h)h_y)_y - K(h)_z = 0 \quad (2)$$

where  $\bar{h}$  is known from the previous time step. Here we are in two space dimensions, and  $y$  is the vertical direction.

Next we discretize the space variable by the finite difference method. In the calculations that we later discuss, we consider a two dimension flow with zero flow through the sides and bottom, and nonzero flow through the top. The finite difference grid is illustrated in Figure 1 where the nine types of boundary cells are indicated. Here there are  $N = 3$  cells in each direction and  $N^2 = 9$  unknowns.





**Figure 1: Finite Difference Grid**

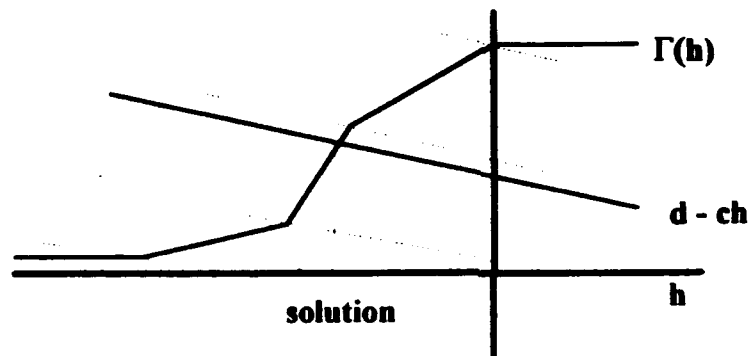
Let  $(i,j)$  denote the location in the finite difference grid. Then the general form of the finite difference equation at this location is

$$d_{i,j} - c_{i,j}h_{i,j} = \Gamma(h_{i,j}) \quad 1 \leq i \leq n_x \text{ and } 1 \leq j \leq n_y. \quad (3)$$

If  $(i,j)$  is an interior node, then

$$\begin{aligned} \Gamma(h_{i,j}) &= \frac{\theta(h_{i,j})}{\Delta x} + \frac{K(h_{i,j})}{\Delta y} \\ c_{i,j} &= (K_{i-1/2,j} + K_{i+1/2,j}) \frac{1}{\Delta x^2} + (K_{i,j-1/2} + K_{i,j+1/2}) \frac{1}{\Delta y^2} \\ d_{i,j} &= \frac{\theta(\bar{h}_{i,j})}{\Delta x} + (K_{i-1/2,j}h_{i-1,j} + K_{i+1/2,j}h_{i+1,j}) \frac{1}{\Delta x^2} \\ &\quad + (K_{i,j-1/2}h_{i,j-1} + K_{i,j+1/2}h_{i,j+1}) \frac{1}{\Delta y^2} \\ &\quad + K(h_{i,j}) \frac{1}{\Delta y}. \end{aligned}$$

In the above equation we used the convention that  $K_{i+1/2,j+1/2}$  is the average of the hydraulic conductivity at the appropriate surrounding nodes. We also will assume that the surrounding nodes are evaluated at a "previous iteration" value. Thus equation (3) is a piecewise linear system as illustrated in Figure 2. Both the piecewise linear approximation of the moisture content and hydraulic conductivity functions are monotone, and so,  $\Gamma$  must also be monotone nondecreasing. Since the term on the left side of equation (3) has negative slope, equation (3) has a solution, and it is unique. In Figure 2 the data is depicted as being continuous, but this need not be the case. Even if  $\Gamma(h)$  has a jump and remains nondecreasing, one can still solve for a unique  $h$  (Silva Neto and White 1993).



**Figure 2: Solution of Equation (3)**

### 3. NONLINEAR SOR ALGORITHM

The following algorithm has evolved from the work by Cryer (1971) for set valued systems of equations that may come from models of the Stefan problem. We apply a variation to the system given in (3). The following variables are used:

$\text{maxit}$  = number of allowable SOR iterations per time step,

$n_x, n_y$  = number of cells in the x and y direction,

$\bar{\omega}, \underline{\omega}$  = overrelaxation (larger than 1.0) and underrelaxation (less than 1.0).

### Nonlinear SOR Algorithm

```
for k = 1,maxit
  for i = 1,nx
    for j = 1,ny
      compute ci,j and di,j as given in (3)
      solve for h in (3) as given in Figure 2
      if h and hi,j are in the same linear phase, then
        
$$h_{i,j} = (1 - \bar{\omega})h_{i,j} + \bar{\omega}h$$

      else
        
$$h_{i,j} = (1 - \underline{\omega})h_{i,j} + \underline{\omega}h$$

      endif
    end loop j
  end loop i
  test for convergence
end loop k.
```

In the computation of the c and d values one must consider the nine types of cells as indicated in Figure 1. For more complicated geometric configurations and for heterogeneous porous media, this will be more complicated. If adjacent cells have different moisture content and hydraulic conductivity data, then one must insert additional nodes between the cells and demand continuity of the flow velocity at the interface (White et al. 1993).

In the solve step one must determine which linear phase the solution is in, and this is done by partitioning the vertical axis as indicated in Figure 2 by the dotted lines that are parallel to the line given by  $d - ch$ . Hence, the solve step has a loop in it which was not indicated above. This hidden loop contains the nonlinear nature of the solve step.

Moreover, if there is a large number of linear phases, then the solve step will become more expensive to compute.

The overrelaxation is used to reduce the number of outer iterations, and the optimal choice will vary with the number of unknowns. The underrelaxation is used to avoid numerical oscillations, and this works well for choices between 0.8 and 0.9. The numerical oscillations are a result of passing from one linear phase to the next linear phase. This deals with the large derivatives of the data by breaking the changes in slopes into a number of smaller changes in slope. We found this to be much more effective than the traditional method of reducing the time step.

We experimented with a number of convergence tests. Finally, we imposed two conditions:

$$(i) \quad \max | \text{new } h - \text{old } h | \leq \varepsilon_1 \quad \text{and}$$

$$(ii) \quad \iint | \text{new } \theta - \text{old } \theta | \leq \varepsilon_2.$$

The first condition is aimed at possible convergence of the pressure at each node. The second condition reflects possible convergence of the total moisture, and it is more of a global test than the first condition.

#### **4. COMPUTATIONS FOR BRINDABELLA LOAM**

The purpose of these computations is to see if our model of Richards' equation will accurately track the movement of moisture through Brindabella loam. We compare our calculations with the observations in White and Broadbridge (1988). In our numerical model we considered a 0.3[m] x 0.3[m] region with boundary conditions as indicated in Figure 1. In the top boundary we used  $R = 0.0165[\text{m/hr}]$ , and the initial pressure was set as  $h = -8.0[\text{m}]$ . The moisture content function was a linear interpolation of the data in Table 1. The hydraulic conductivity data indicates a very increasing and concave up function; consequently, linear interpolation of the data would generate large errors. In the

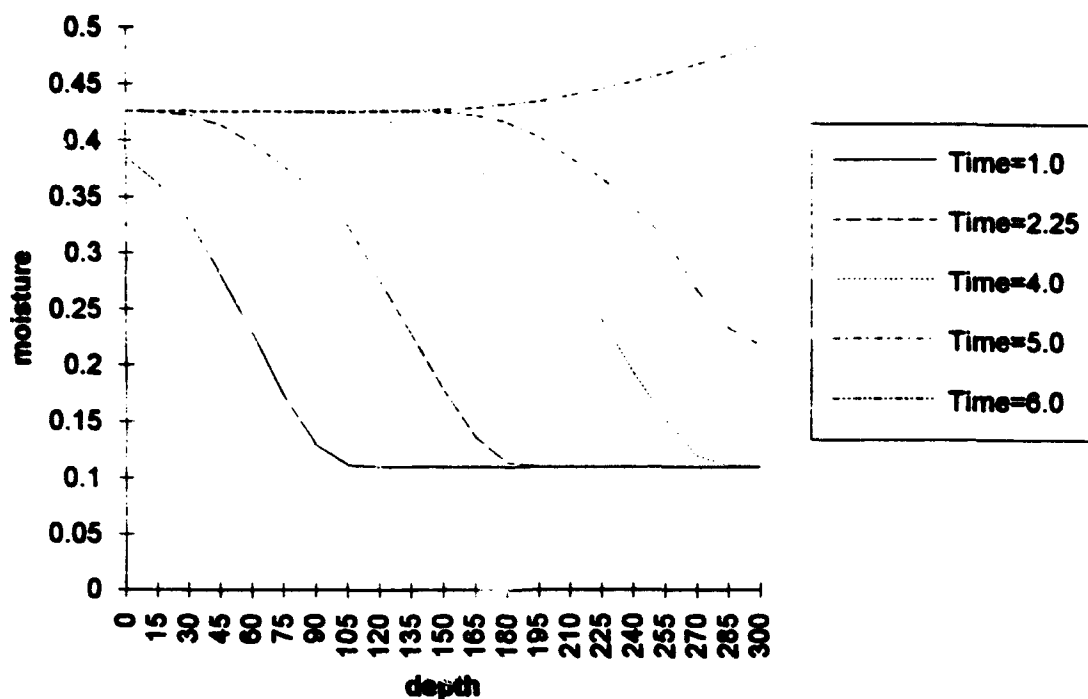
calculations presented in Figure 3a we used a linear interpolation of the modified data for hydraulic conductivity in Table 1; we reduced the interior values by 50 percent and kept the two end values at 0.0 and 0.118.

In our computations we set the following parameters at

$$\begin{array}{ll} \Delta t = 0.125[\text{hr}] & \Delta x = 0.015[\text{m}] \\ N = 20[\text{cells in each dir.}] & R = 0.0165[\text{m/hr}] \\ \omega = 0.9 & \omega = 1.4 \\ \varepsilon_1 = 10^{-4} & \varepsilon_2 = 10^{-8} \end{array}$$

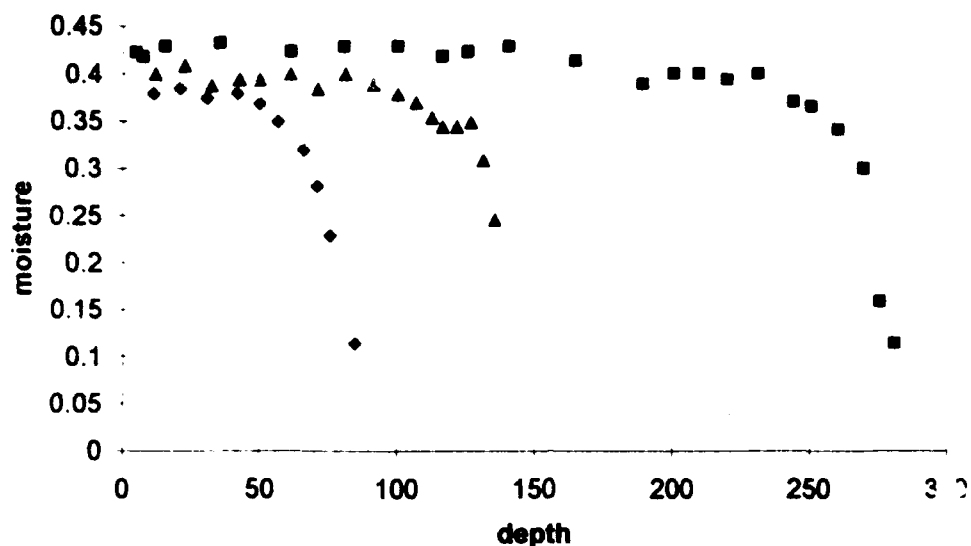
Convergence was usually attained in 20-40 iterations. If no underrelaxation was used, then numerical oscillations would occur about once every 20 time steps.

During the initial times, the hydraulic conductivity is small, and Richards' equation is dominated by wave like properties. As time progresses the hydraulic conductivity increases so that Richards' equation is dominated by diffusion. At time 6.0[hr] the steady state solution has essentially been reached. At the bottom ( $y = 300[\text{mm}]$ ) the porous media is at saturation ( $\theta = 0.485$ ). At the top ( $y=0[\text{mm}]$ ) the porous media has pressure such that  $K(h) = R$  ( $\theta(h) = 0.426$ ). At this time the diffusion force is equal to the gravitational force; hence, no more moisture can enter the porous media.



**Figure 3a: Computed Moisture for Variable Times**

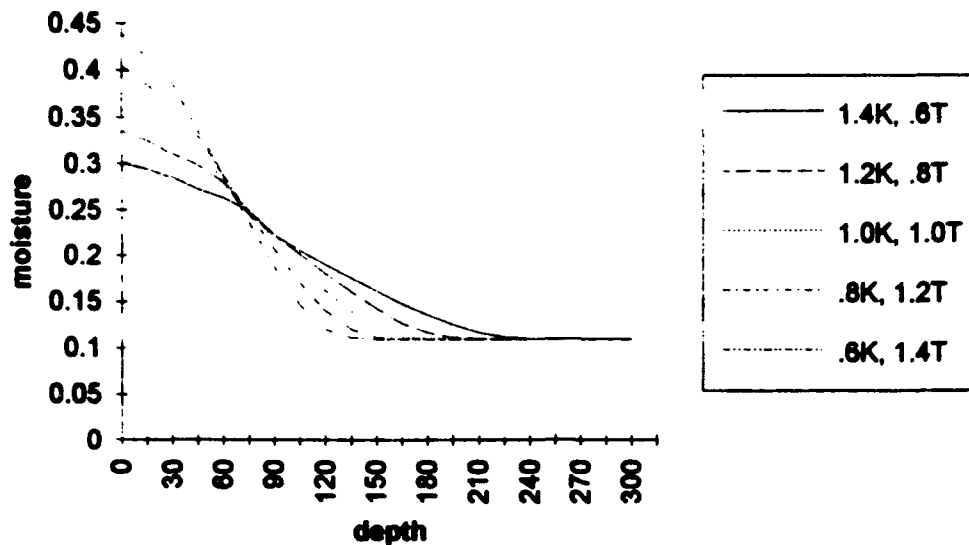
The observed moistures are indicated by discrete points in Figure 3b. These were for the times of 1.0 (diamonds), 2.25 (triangles) and 5.0 hours (squares). The computed values give good agreement with the observed values. The computed moisture content curves are somewhat more smoothed than the observed moisture content data. This may be attributed to large hydraulic conductivity data; if one reduces the hydraulic conductivity for smaller pressures, then a sharp front can be calculated to match the observed moisture content data.



**Figure 3b: Observed Moisture Data for Variable Times**

## 5. COMPUTATIONS WITH UNCERTAIN DATA

This section contains an analysis of the moisture as a function of the empirical data of the moisture content and hydraulic conductivity functions. In practice much of this data is not precisely known, and therefore, the effects upon the computations from any model will have some uncertain aspects. In our numerical experiments we decreased  $K$  and the computed moisture at the top increased. We also increased  $\theta$  and the computed moisture at the top increased. In the computations indicated in Figure 4, for time equal to 1.25[hr], we decreased  $K$  and increased  $\theta$ , and the largest computed moisture content at the top was the result. Here we kept the data at the end points of Table 1 fixed and varied the interior data by increments of 20 percent. In all computations the computed moisture content at the top increased while the computed moisture content at the bottom decreased. This happens because the sides and bottom do not permit flow through them, and the total moisture must remain constant.



**Figure 4: Moisture and Variable Data**

In order to gain some insight on this, it is instructive to examine the finite difference equation at the top region. In the case of the top center nodes,  $d$  has the form

$$d = \frac{2R}{\Delta y} - \frac{K(h_{i,j-1})}{\Delta y} + \bar{d} \quad \text{where } \bar{d} \text{ has form similar to that given in (3).}$$

Figure 5 shows that if  $\Gamma$  increases, then the solution of (3) will decrease. Also, if  $d$  decreases, then the solution of (3) will decrease. Therefore, if both  $\Gamma$  increases and  $d$  decreases, then the solution of (3) will decrease. If  $K$  decreases, then  $d$  will increase and  $\Gamma$  will decrease. However, if both  $K$  decreases, and  $\theta$  increases enough, then  $\Gamma$  will increase. For our choices of  $\Delta t$  and  $\Delta y$  this is the case. Of course, this is just an analysis at one grid point, and the argument requires much more careful discussion.



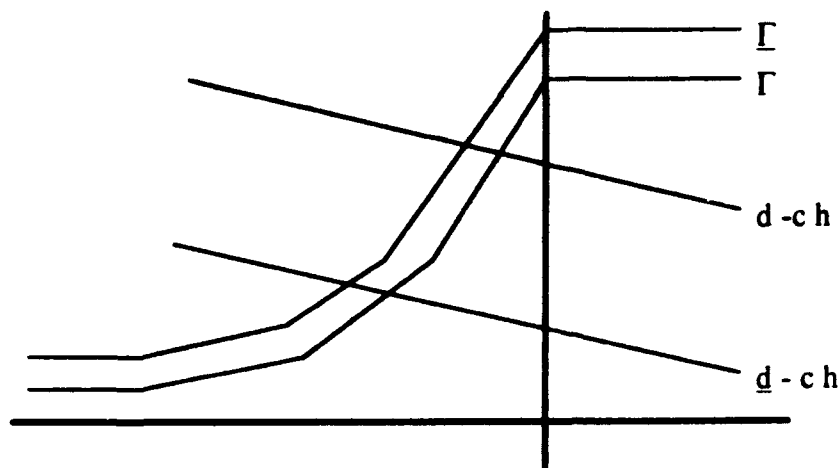


Figure 5: Moisture and Variable Data Analysis

## 6. COMPUTATIONS USING MULTIPROCESSORS

In the computations reported in this section we tried to implement the above algorithm on a single CPU with vectorization on a Cray Y-MP, the Alliant FX-40 with two vectorized CPUs, and the Kendall Square Research KSR1 with up to 16 CPUs and no vectorization. In the calculations in Silva Neto and White (1993) the vectorization methods did not seem to work well. These attempts involved reordering the nodes by the red-black order (checker board order). This method also did not work well for our current problem. The reason for this is that the inner most loop has computations which are too complicated to effectively be done on a vector pipeline.

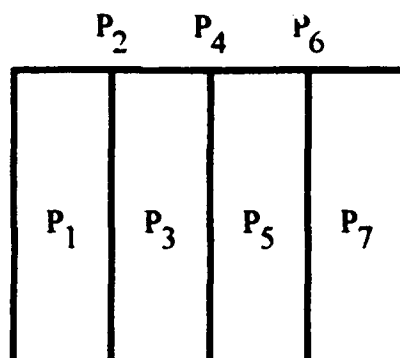
The multiprocessing approach with domain decomposition reordering (White 1987, or Ortega 1988) was much more promising. This reordering is depicted in Figure 5 where  $L = 4$  (the number of larger blocks of nodes) and the classical order of the blocks of grid points is

$$P_1, P_2, P_3, P_4, P_5, P_6, P_7$$

The domain decomposition order lists all the smaller interior boundary blocks first (even number blocks in Figure 6) and is

$$P_2, P_4, P_6, P_1, P_3, P_5, P_7$$

The idea behind this reordering is to take advantage of the 5-point finite difference pattern. Once the calculations in the even blocks have been done, then the calculations in the large odd blocks are independent of one another.



**Figure 6: Domain Decomposition Order**

#### Nonlinear SOR Algorithm: Domain Decomposition

```

for k = 1,maxit
    concurrently do SOR over the even blocks
    update
    concurrently do SOR over the odd blocks
    test for convergence
end loop k.

```

In our calculations we used the Kendall Square Research multiprocessing computer, KSR1, which is operated by the North Carolina Supercomputing Center. The KSR1 multiprocessing computer has three parallel constructs that can be used in FORTRAN code: *tile*, *parallel section* and *parallel region*. Tile is used to partition loops and is very effective for simple computations such as matrix multiplications. Parallel section can be used to concurrently execute different code segments. We used parallel region which duplicates a code segment and uses different data streams. In our

computations we controlled the number of processors by using a *team of processors* that are assigned at the beginning of the code and are used to reduce parallel overhead.

Table 2 shows the speedup and efficiency for a variety of  $L$  (the number of large blocks) and  $N$  (the number of cells in each direction). These quantities are defined as follows:

$$S_L = (\text{CPU time using one block})/(\text{CPU time using } L \text{ large blocks}) \text{ and}$$

$$E_L = S_L/L.$$

In the first four rows  $N$  varies and  $L$  is fixed. We see increased speedup and efficiency as  $N$  increases. This is a result of decreased parallel overhead. In the last four rows  $N$  is fixed, and  $L$  is increased. Here the speedup increases, but the efficiency decreases. Of course, if there are many larger blocks ( $L$ ) and the number of cells in each direction ( $N$ ) remains the same, then the relative size of the larger block to the smaller block decreases. This partially accounts for the decreased efficiency. These calculations did not attempt to make the most efficient use of the FORTRAN language, or the most efficient use of the KSR1 computer's architecture.

**Table 2: Speedup and Efficiency**

$N$	$L$	$S_L$	$E_L$
20	2	1.56	0.78
40	2	1.68	0.84
80	2	1.75	0.88
160	2	1.78	0.89
160	4	3.16	0.79
160	8	5.09	0.64
160	16	7.29	0.46

## 7. CONCLUSIONS

Richards' equation was approximated by the finite difference method, and the empirical data for the moisture content and the hydraulic conductivity were approximated by piecewise linear functions. The resulting nonlinear algebraic system was solved by a variation of the nonlinear SOR iterative method. Good convergence properties were observed for three types of calculations which were chosen to demonstrate the feasibility of realistic numerical simulations using this method. These included an accurate simulation of fluid flow in Brindabella loam, a sensitivity analysis of the computed solution upon the empirical data, and the use of multiprocessing computers via domain decomposition methods.

In the above calculations the empirical data did not have a space dependence. However, in White et al. (1993) we illustrated for a steady state and one space dimension version of Richards' equation that the compact volume method, in place of the finite difference method, could be effectively used for such heterogeneous problems. The compact volume method can be viewed as an enhanced finite difference method where additional nodes are inserted at the cell interface and additional equations are generated by requiring continuity of the fluid velocity at these interfaces. This may be done for all cell interfaces or for just those cells where the empirical functions change with respect to the space variable. We expect the methods of this paper to generalize via the compact volume method to the more complicated heterogeneous case.

## REFERENCES

- Bear, J. (1972), *Dynamics of Fluids in Porous Media*, Dover, New York.
- Clothier, B.E., White, I., and Hamilton, G.J. (1981), "Constant rate rainfall infiltration: field experiments," *Soil Sci. Soc. Am. J.*, vol. 45, pp. 245-249.
- Cryer, C.W. (1971), "The solution of a quadratic programming problem using systematic overrelaxation," *SIAM J. Control*, vol. 9, no. 3, pp.385-392.
- Feng, J., (1993), *Modeling of Chemical Transport from Agricultural Waste Lagoons*, Ph.D. Dissertation, North Carolina State University.
- Freeze, R.A. and Cherry, J.A. (1979), *Groundwater*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- Kaviany, M. (1991), *Principles of Heat Transfer in Porous Media*, Mechanical Engineering Series, Springer-Verlag, New York.
- Kimura, S. (1989), "Transient forced convection heat transfer from a circular cylinder in a saturated porous medium," *Int. J. Heat and Mass Transfer*, vol. 32, no. 1, pp.192-195.
- Kimura, S. (1989a), "Transient forced natural convection heat transfer about a vertical cylinder in a porous medium," *Int. J. Heat and Mass Transfer*, vol.32, no. 3, pp.617-620.
- Mark, S.M., (1992), "Groundwater modeling: a vital tool for water resource management," *Colorado School of Mines Quarterly Review*, vol. 92, no. 4, pp.9-12.
- Mark, S.M., (1993), "Groundwater modeling: a vital tool for water resource management, Part II," *Colorado School of Mines Quarterly Review*, vol. 93, no. 1, pp.1-5.
- Muralidhar, K. (1990), "Flow and transport in single rock fractures," *J. Fluid Mech.*, vol. 215, pp.481-502.
- Muralidhar, K. (1993), "Near-field solution for heat and mass transfer from buried nuclear waste canisters," *Int. J. Heat and Mass Transfer*, vol. 36, no. 10, pp.2665-2674.
- Nield, D.A. and Bejan, A. (1992), *Convection in Porous Media*, Springer-Verlag, New York.

Ortega, J.M. (1988), *Introduction to Parallel and Vector Solution of Linear Systems*, Plenum Press, New York.

Paniconi, C., Aldama, A.A. and Wood, E.F. (1991), "Numerical evaluation of iterative and noniterative methods for the solution of the nonlinear Richards' equation," *Water Resources Research*, vol. 27, no. 6, pp.1147-1163.

Philip, J.R. (1969), "Theory of infiltration," *Adv. Hydrosci.*, vol. 5, pp.215-296.

Rae, J., Robinson, P.C. and Wickens, L.M. (1983), "Coupled heat and groundwater flow in porous rock," *Numerical Methods in Heat Transfer, Volume II*, edited by Lewis, R.W., Morgan, K. and Schrefler, B.A., Wiley, New York.

Richards, L.A. (1931), "Capillary conduction of liquids through porous media," *Physics*, vol.1, pp.318-333.

Silva Neto, J.A. and White, R.E. (1993), "Numerical solution of Richards' equation," to appear in the proceedings of the 12<sup>th</sup> Brazilian Congress of Mechanical Engineering.

van Genuchten, M.T. and Nielsen, D.R. (1985), "On describing and predicting the hydraulic properties of unsaturated soils," *Ann. Geophys.*, vol. 3, no. 5, pp.615-628.

White, R.E. (1985), *An Introduction to the Finite Element Method with Applications to Nonlinear Problems*, Wiley, New York.

White, R.E. (1987), "Multisplitting and parallel iterative methods," *J. Comp. Meth. in Appl. Mech. and Eng.*, vol. 64, pp.567-577.

White, I. and Broadbridge, P. (1988), "Constant rate rainfall infiltration: a vertical model. 2. Applications of solutions," *Water Resouces Research*, vol. 24. no.1, pp.155-162.

White, R.E., Borah, B.N. and Kyrillidis, A.J. (1993), "Heterogeneous diffusion and the compact volume method," submitted.

# Explicit formal solution to generalized Kolmogorov equation \*

**Shing-Tung Yau**  
Department of Mathematics  
Harvard University  
Cambridge, MA  
02138  
U.S.A.

**Stephen S.-T. Yau**  
Control and Information Laboratory  
Department of Mathematics  
Statistics and Computer Sciences  
University of Illinois at Chicago  
851 South Morgan Street  
Chicago, IL. 60607-7045  
U.S.A.

U32790@UICVM.BITNET

## Abstract

It is well known that Kolmogorov equation is a fundamental equation in Applied Science, especially in Electrical Engineering. Our original motivation is to solve the Duncan-Mortensen-Zakai equation in nonlinear filtering theory. If the observation  $h(x)$  is a constant in Duncan-Mortensen-Zakai (DMZ) equation, then it becomes the famous Kolmogorov equation. If we treat  $h(x)$  as a function again, then the resulting equation is called generalized Kolmogorov equation. In this paper, we write down the formal solution of this generalized Kolmogorov equation in a closed form. We shall report the convergent solution in the subsequent paper.

## § 1. Introduction

In the sixties and early seventies, the basic approach to non-linear filtering theory was via "innovations methods", originally proposed by Kailath in 1967 and subsequently rigorously developed by Fujisaki, Kallianpur and Kunita ([F-K-K] 1972) in their seminal paper. As pointed out by Mitter, the difficulty with this approach is that the innovations process is not, in general, explicitly computable (except in the well-known Kalman-Bucy case). The idea of using estimation algebras to construct finite dimensional nonlinear filters was first proposed in the early eighties by Brockett and Clark [Br - Cl], Brockett [Br] and Mitter [Mi]. The motivation came from the Wei-Norman approach [We - No] of using Lie algebraic ideas to solve time variant linear differential equations. The extension of Wei-Norman's approach to the non-linear filtering problem is much more complicated. Instead of an ordinary differential equation, we have to solve the Duncan-Mortensen-Zakai (DMZ) equation, which is a stochastic partial differential equation. By working on the robust form of the DMZ equation we can reduce the complexity of the problem to that of solving a time variant partial differential equation. Wong in [Wo1] constructed some new finite dimensional estimation algebras and the Wei-Norman approach to synthesize finite dimensional filters. However, the systems considered in [Wo1] are very specific and the question whether the Wei-Norman approach works for a general system with finite dimensional estimation algebra remains open.

---

\* Research supported by Army Grant DAAH-04-93G-0006

Recently, Tam, Wong and the second author [T-W-Y] have examined the properties of finite dimensional estimation algebras and the Wei-Norman approach in detail. There a class of filtering systems having the property that the drift term,  $f$ , of the state evolution equation is a gradient vector field was considered. In [Wo2], the concept of  $\Omega$  is introduced, which is defined as the  $n \times n$  matrix whose  $(i,j)$ -entry is  $\frac{\partial f_i}{\partial x_j} - \frac{\partial f_j}{\partial x_i}$ . In view of Poincare lemma,  $f$  is a gradient vector field if and only if  $\Omega = 0$ . More recently, the second named author [Ya] considered a more general class of filtering systems having the property that  $\frac{\partial f_i}{\partial x_j} - \frac{\partial f_j}{\partial x_i}$  are constant for all  $i, j$  i.e.  $\Omega$  is a skew constant matrix. These include Kalman-Bucy filtering systems and Bene's filtering systems as special cases and finite dimensional filters were constructed explicitly. From Lie algebraic point of view, Chen, Chiou, Leung and the second named author [Ch-Ya] [C-L-Y] have shown that these are most general finite dimensional filters at least for dimension of state space less than four. In many senses, the Lie algebraic viewpoint has been remarkably successful and the recent work has given us a deeper understanding of the DMZ equation which was essential for progress in non-linear filtering, as well as in stochastic control.

However, we should notice that the Wei-Norman approach only reduces the DMZ equation to a finite system of ordinary differential equations and the following generalized Kolmogorov equation

$$(1.0) \quad \frac{\partial u}{\partial t}(t, x) = \left\{ \frac{1}{2} \sum_{i=1}^n \left( \frac{\partial}{\partial x_i} - f_i(x) \right)^2 - \frac{1}{2} \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \sum_{i=1}^n f_i^2(x) + \sum_{i=1}^m h_i^2(x) \right) \right\} u(t, x)$$

$$u(0, x) = \sigma_0$$

It is the purpose of this paper to give a closed form formal solution of the above equation.

## § 2. Basic concepts

The filtering problem considered here is based on the following signal observation model :

$$(2.0) \quad \begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t) & x(0) = x_0, \\ dy(t) = h(x(t))dt + dw(t) & y(0) = 0, \end{cases}$$

in which  $x, v, y$ , and  $w$ , are respectively,  $R^n, R^p, R^m$  and  $R^m$  valued processes, and  $v$  and  $w$  have components which are independent, standard Brownian processes. We further assume that  $n = p$ ;  $f, h$  are  $C^\infty$  smooth, and that  $g$  is an orthogonal matrix. We will refer to  $x(t)$  as the state of the system at time  $t$  and to  $y(t)$  as the observation at time  $t$ .

Let  $\rho(t, x)$  denotes the conditional probability density of the state given the observation  $y(s) : 0 \leq s \leq t$ . It is well known (see [Da-Ma] for example) that  $\rho(t, x)$  is given by normalizing a function,  $\sigma(t, x)$ , which satisfies the following Duncan-Mortensen-Zakai equation:

$$(2.1) \quad d\sigma(t, x) = L_0\sigma(t, x)dt + \sum_{i=1}^m L_i\sigma(t, x)dy_i(t), \quad \sigma(0, x) = \sigma_0,$$



where  $L_0 = \frac{1}{2} \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^n f_i \frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} - \frac{1}{2} \sum_{i=1}^m h_i^2$  and for  $i = 1, \dots, m$ ,  $L_i$  is the zero degree differential operator of multiplication by  $h_i$ .  $\sigma_0$  is the probability density of the initial point,  $x_0$ .

Equation (2.1) is a stochastic partial differential equation. In real applications, we are interested in constructing robust state estimators from observed sample paths with some property of robustness. Davis in [Da] studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new unnormalized density

$$u(t, x) = \exp\left(-\sum_{i=1}^m h_i(x) y_i(t)\right) \sigma(t, x)$$

It is easy to show that  $u(t, x)$  satisfies the following time varying partial differential equation (2.2)

$$\begin{aligned} \frac{\partial u}{\partial t}(t, x) &= L_0 u(t, x) + \sum_{i=1}^m y_i(t) [L_0, L_i] u(t, x) + \frac{1}{2} \sum_{i,j=1}^m y_i(t) y_j(t) [[L_0, L_i], L_j] u(t, x), \\ u(0, x) &= \sigma_0 \end{aligned}$$

where  $[\cdot, \cdot]$  is the Lie bracket defined as:

**Definition :** If  $X$  and  $Y$  are differential operators, the Lie bracket of  $X$  and  $Y$ ,  $[X, Y]$ , is defined by

$$[X, Y]\xi = X(Y\xi) - Y(X\xi)$$

for any  $C^\infty$  function  $\xi$ .

In §3, we shall write down the formal solution of (2.2) explicitly in closed form.

### § 3. Formal solution to generalized Kolmogorov equation

The purpose of this section is to write down a formal solution of the time varying differential equation (2.2).

**Lemma 1** Equation (2.2) is equivalent to the following equation.

(3.0)

$$\begin{aligned} \frac{\partial u}{\partial t}(t, x) &= \left\{ \frac{1}{2} \sum_{i=1}^n \left[ \frac{\partial}{\partial x_i} - [f_i(x) - \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_i}(x)] \right]^2 \right. \\ &\quad \left. - \frac{1}{2} \left[ \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \sum_{i=1}^n f_i^2(x) + \sum_{i=1}^m h_i^2(x) \right] \right\} u(t, x) \\ u(0, x) &= \sigma_0(x) \end{aligned}$$

# **Proof**

$$\begin{aligned}
 [L_0, h_i(x)] &= \frac{1}{2} \left[ \sum_{j=1}^n \left( \frac{\partial}{\partial x_j} - f_j(x) \right)^2 - \left( \sum_{j=1}^n \frac{\partial f_j}{\partial x_j}(x) + \sum_{j=1}^n f_j^2(x) + \sum_{j=1}^m h_j^2(x) \right), h_i(x) \right] \\
 &= \frac{1}{2} \left[ \sum_{j=1}^n \left( \frac{\partial}{\partial x_j} - f_j(x) \right)^2, h_i(x) \right] \\
 &= \frac{1}{2} \sum_{j=1}^n \frac{\partial^2 h_i}{\partial x_j^2}(x) + \sum_{j=1}^n \frac{\partial h_i}{\partial x_j}(x) \frac{\partial}{\partial x_j} - \sum_{j=1}^n f_j(x) \frac{\partial h_i}{\partial x_j}(x) \\
 &= \sum_{j=1}^n \frac{\partial h_i}{\partial x_j}(x) \frac{\partial}{\partial x_j} + \frac{1}{2} \sum_{j=1}^n \frac{\partial^2 h_i}{\partial x_j^2}(x) - \sum_{j=1}^n f_j(x) \frac{\partial h_i}{\partial x_j}(x) \\
 [L_0, h_i], h_j &= \sum_{k=1}^n \left[ \frac{\partial h_i}{\partial x_k}(x) \right] \left[ \frac{\partial h_j}{\partial x_k}(x) \right]
 \end{aligned}$$

$$\begin{aligned}
 L_0 + \sum_{i=1}^m y_i(t) [L_0, L_i] + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i(t) y_j(t) [[L_0, L_i], L_j] \\
 = \frac{1}{2} \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^n f_i(x) \frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) - \frac{1}{2} \sum_{i=1}^m h_i^2(x) + \sum_{i=1}^m \sum_{j=1}^n y_i(t) \frac{\partial h_i}{\partial x_j}(x) \frac{\partial}{\partial x_j} \\
 + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n y_i(t) \frac{\partial^2 h_i}{\partial x_j^2}(x) - \sum_{i=1}^m \sum_{j=1}^n y_i(t) f_j(x) \frac{\partial h_i}{\partial x_j}(x) \\
 + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n y_i(t) y_j(t) \left[ \frac{\partial h_i}{\partial x_k}(x) \right] \left[ \frac{\partial h_j}{\partial x_k}(x) \right] \\
 = \frac{1}{2} \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} + \sum_{i=1}^n \left[ -f_i(x) + \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_i}(x) \right] \frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) - \frac{1}{2} \sum_{i=1}^m h_i^2(x) \\
 + \frac{1}{2} \sum_{i=1}^m y_i(t) \Delta h_i(x) - \sum_{i=1}^m \sum_{j=1}^n y_i(t) f_j(x) \frac{\partial h_i}{\partial x_j}(x) \\
 + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i(t) y_j(t) \sum_{k=1}^n \left[ \frac{\partial h_i}{\partial x_k}(x) \right] \left[ \frac{\partial h_j}{\partial x_k}(x) \right] \\
 = \frac{1}{2} \sum_{i=1}^n \left[ \frac{\partial}{\partial x_i} - \left[ f_i(x) - \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_i}(x) \right] \right]^2 + \frac{1}{2} \sum_{i=1}^n \left[ \frac{\partial f_i}{\partial x_i}(x) - \sum_{j=1}^m y_j(t) \frac{\partial^2 h_j}{\partial x_i^2}(x) \right] \\
 - \frac{1}{2} \sum_{i=1}^n \left[ f_i(x) - \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_j}(x) \right]^2 - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) - \frac{1}{2} \sum_{i=1}^m h_i^2(x) + \frac{1}{2} \sum_{i=1}^m y_i(t) \Delta h_i(x) \\
 - \sum_{i=1}^m \sum_{j=1}^n y_i(t) f_j(x) \frac{\partial h_i}{\partial x_j}(x) + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i(t) y_j(t) \sum_{k=1}^n \left[ \frac{\partial h_i}{\partial x_k}(x) \right] \left[ \frac{\partial h_j}{\partial x_k}(x) \right]
 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \sum_{i=1}^n \left[ \frac{\partial}{\partial x_i} - \left[ f_i(x) - \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_i}(x) \right] \right]^2 - \frac{1}{2} \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) - \frac{1}{2} \sum_{i=1}^n f_i^2(x) \\
&\quad + \sum_{i=1}^n \sum_{j=1}^m f_i(x) y_j(t) \frac{\partial h_j}{\partial x_i}(x) - \frac{1}{2} \sum_{i=1}^n \left[ \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_i}(x) \right]^2 - \frac{1}{2} \sum_{i=1}^m h_i^2(x) \\
&\quad - \sum_{i=1}^m \sum_{j=1}^n y_i(t) f_j(x) \frac{\partial h_i}{\partial x_j}(x) + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i(t) y_j(t) \sum_{k=1}^n \left[ \frac{\partial h_i}{\partial x_k}(x) \right] \left[ \frac{\partial h_j}{\partial x_k}(x) \right] \\
&= \frac{1}{2} \sum_{i=1}^n \left[ \frac{\partial}{\partial x_i} - \left[ f_i(x) - \sum_{j=1}^m y_j(t) \frac{\partial h_j}{\partial x_i}(x) \right] \right]^2 - \frac{1}{2} \left[ \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \sum_{i=1}^n f_i^2(x) + \sum_{i=1}^m h_i^2(x) \right]
\end{aligned}$$

Q.E.D.

We observe that if  $h_i(x)$ ,  $1 \leq i \leq m$  are constants in the robust DMZ equation (3.0), then (3.0) becomes the Kolmogorov equation

$$\begin{aligned}
\frac{\partial u}{\partial t}(t, x) &= \left\{ \frac{1}{2} \sum_{i=1}^n \left( \frac{\partial}{\partial x_i} - f_i(x) \right)^2 - \frac{1}{2} \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \sum_{i=1}^n f_i^2(x) + \sum_{i=1}^m h_i^2(x) \right) \right\} u(t, x) \\
u(0, x) &= \sigma_0
\end{aligned}$$

In generalized Kolmogorov equation (1.0) is the above equation which we let  $h_i$  depend on  $x$  again.

**Theorem 2** The equation (1.0) has a formal asymptotic solution on  $R^n$ . In fact, the solution is of the following form

$$(3.1) \quad u(t, x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-n/2} \exp\left(-\sum_{j=1}^n (x_j - \xi_j)^2 / 2t\right) b(t, x, \xi) \sigma_0(\xi) d\xi_1 \cdots d\xi_n$$

where  $b(t, x, \xi) = \sum_{k=0}^{\infty} a_k(x, \xi) t^k$ .

Here  $a_k(x, \xi)$  are described explicitly as follows. Let

$$(3.2) \quad a(x, \xi) = \int_0^1 \sum_{i=1}^n (x_i - \xi_i) f_i[\xi + t(x - \xi)] dt$$

Then

$$(3.3) \quad a_0(x, \xi) = e^{a(x, \xi)}$$

Suppose that  $a_{k-1}(x, \xi)$  is given. Let

$$\begin{aligned}
(3.4) \quad g_k(x, \xi) &= \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) - \sum_{i=1}^n f_i(x) \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) \\
&\quad - \frac{1}{2} \left( \sum_{i=1}^m h_i^2 \right) a_{k-1}(x, \xi) - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) a_{k-1}(x, \xi)
\end{aligned}$$

Then, for  $k \geq 1$ ,

$$a_k(x, \xi) = e^{a(x, \xi)} \int_0^1 t^{k-1} e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi + t(x-\xi), \xi) dt$$

**Proof** We shall prove that (3.1) is a formal solution of (1.0). Putting (3.1) into (1.0) we have

$$\begin{aligned} L.H.S. \text{ of (1.0)} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n (x_j - \xi_j)^2 / 2t\right) \\ &\quad \cdot \left[ \sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t} b(t, x, \xi) - \frac{n}{2} b(t, x, \xi) + t \frac{\partial b}{\partial t}(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n \end{aligned}$$

On the other hand,

$$\begin{aligned} \left[ \frac{\partial}{\partial x_i} - f_i(x) \right] u(t, x) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \\ &\quad \cdot \left[ -(x_i - \xi_i) b(t, x, \xi) + t \frac{\partial b}{\partial x_i}(t, x, \xi) - t f_i(x) b(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n \end{aligned}$$

$$\begin{aligned} &\left[ \frac{\partial}{\partial x_i} - f_i(x) \right]^2 u(t, x) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \frac{-(x_i - \xi_i)}{t} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \left[ -(x_i - \xi_i) b(t, x, \xi) \right. \\ &\quad \left. + t \frac{\partial b}{\partial x_i}(t, x, \xi) - t f_i(x) b(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n \\ &\quad + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \left[ -b(t, x, \xi) - (x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) \right. \\ &\quad \left. + t \frac{\partial^2 b}{\partial x_i^2}(t, x, \xi) - t \frac{\partial f_i}{\partial x_i}(x) b(t, x, \xi) - t f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n \\ &\quad + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \left[ (x_i - \xi_i) f_i(x) b(t, x, \xi) \right. \\ &\quad \left. - t f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) + t f_i^2(x) b(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \left[ \frac{(x_i - \xi_i)^2}{t} b(t, x, \xi) \right. \\
&\quad - (x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) + (x_i - \xi_i) f_i(x) b(t, x, \xi) - b(t, x, \xi) \\
&\quad - (x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) + t \frac{\partial^2 b}{\partial x_i^2}(t, x, \xi) - t \frac{\partial f_i}{\partial x_i}(x) b(t, x, \xi) \\
&\quad - t f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) + (x_i - \xi_i) f_i(x) b(t, x, \xi) - t f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) \\
&\quad \left. + t f_i^2(x) b(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \left[ \frac{(x_i - \xi_i)^2}{t} b(t, x, \xi) \right. \\
&\quad - 2(x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) + 2(x_i - \xi_i) f_i(x) b(t, x, \xi) \\
&\quad - b(t, x, \xi) + t \frac{\partial^2 b}{\partial x_i^2}(t, x, \xi) - t \frac{\partial f_i}{\partial x_i}(x) b(t, x, \xi) - 2t f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) \\
&\quad \left. + t f_i^2(x) b(t, x, \xi) \right] \sigma_0(\xi) d\xi_1 \cdots d\xi_n
\end{aligned}$$

R.H.S. of (1.0)

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-\frac{n}{2}-1} \exp\left(-\sum_{j=1}^n \frac{(x_j - \xi_j)^2}{2t}\right) \left[ \frac{1}{2t} \sum_{i=1}^n (x_i - \xi_i)^2 b(t, x, \xi) \right. \\
&\quad - \sum_{i=1}^n (x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) + \sum_{i=1}^n (x_i - \xi_i) f_i(x) b(t, x, \xi) - \frac{n}{2} b(t, x, \xi) \\
&\quad + \frac{t}{2} \sum_{i=1}^n \frac{\partial^2 b}{\partial x_i^2}(t, x, \xi) - \frac{t}{2} \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) b(t, x, \xi) - t \sum_{i=1}^n f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) \\
&\quad + \frac{t}{2} \sum_{i=1}^n f_i^2(x) b(t, x, \xi) - \frac{t}{2} \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \sum_{i=1}^n f_i^2(x) + \sum_{i=1}^m h_i^2(x) \right) b(t, x, \xi) \Big] \\
&\quad \cdot \sigma_0(\xi) d\xi_1 \cdots d\xi_n
\end{aligned}$$

It follows that (3.1) is a solution of (1.0) if

$$\begin{aligned}
t \frac{\partial b}{\partial t}(t, x, \xi) &= - \sum_{i=1}^n (x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) + \sum_{i=1}^n (x_i - \xi_i) f_i(x) b(t, x, \xi) \\
&\quad + \frac{t}{2} \sum_{i=1}^n \frac{\partial^2 b}{\partial x_i^2}(t, x, \xi) - \frac{t}{2} \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) b(t, x, \xi) - t \sum_{i=1}^n f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) \\
&\quad + \frac{t}{2} \sum_{i=1}^n f_i^2(x) b(t, x, \xi) - \frac{t}{2} \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \sum_{i=1}^n f_i^2(x) + \sum_{i=1}^m h_i^2(x) \right) b(t, x, \xi)
\end{aligned}$$

i.e. if

$$\begin{aligned}
 t \frac{\partial b}{\partial t}(t, x, \xi) = & - \sum_{i=1}^n (x_i - \xi_i) \frac{\partial b}{\partial x_i}(t, x, \xi) + \sum_{i=1}^n (x_i - \xi_i) f_i(x) b(t, x, \xi) \\
 (3.5) \quad & + \frac{t}{2} \sum_{i=1}^n \frac{\partial^2 b}{\partial x_i^2}(t, x, \xi) - t \sum_{i=1}^n f_i(x) \frac{\partial b}{\partial x_i}(t, x, \xi) \\
 & + \frac{t}{2} \left[ -2 \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) - \sum_{i=1}^m h_i^2(x) \right] b(t, x, \xi)
 \end{aligned}$$

Put  $b(t, x, \xi) = \sum_{k=0}^{\infty} a_k(x, \xi) t^k$  in (3.5). We have

*R.H.S. of (3.5)*

$$\begin{aligned}
 & = - \sum_{i=1}^n \sum_{k=0}^{\infty} (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) t^k + \sum_{i=1}^n \sum_{k=0}^{\infty} (x_i - \xi_i) f_i(x) a_k(x, \xi) t^k \\
 & + \sum_{i=1}^n \sum_{k=0}^{\infty} \frac{1}{2} \frac{\partial^2 a_k}{\partial x_i^2}(x, \xi) t^{k+1} - \sum_{i=1}^n \sum_{k=0}^{\infty} f_i(x) \frac{\partial a_k}{\partial x_i}(x, \xi) t^{k+1} \\
 & - \sum_{k=0}^{\infty} \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^m h_i^2(x) \right) a_k(x, \xi) t^{k+1} \\
 & = - \sum_{k=0}^{\infty} \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) t^k + \sum_{k=0}^{\infty} \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_k(x, \xi) t^k \\
 & + \sum_{k=1}^{\infty} \sum_{i=1}^n \frac{1}{2} \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) t^k - \sum_{k=1}^{\infty} \sum_{i=1}^n f_i(x) \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) t^k \\
 & - \sum_{k=1}^{\infty} \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^m h_i^2(x) \right) a_{k-1}(x, \xi) t^k \\
 & = - \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_0}{\partial x_i}(x, \xi) + \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_0(x, \xi) \\
 & + \sum_{k=1}^{\infty} \left[ - \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) + \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_k(x, \xi) \right. \\
 & + \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) - \sum_{i=1}^n f_i(x) \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) \\
 & \left. - \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^m h_i^2(x) \right) a_{k-1}(x, \xi) \right] t^k
 \end{aligned}$$

$$\text{L.H.S. of (3.5)} = \sum_{k=1}^{\infty} k a_k(x, \xi) t^k$$

Therefore (3.1) is a solution of (1.0) if the following (3.6) and (3.7) are satisfied

$$(3.6) \quad \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_0}{\partial x_i}(x, \xi) = \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_0(x, \xi)$$

(3.7) For  $k \geq 1$

$$\begin{aligned} & \left( k - \sum_{i=1}^n (x_i - \xi_i) f_i(x) \right) a_k(x, \xi) + \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) \\ &= \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) - \sum_{i=1}^n f_i(x) \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) - \left( \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2} \sum_{i=1}^n h_i^2(x) \right) a_{k-1}(x, \xi) \end{aligned}$$

Differentiate (3.2) with respect to  $x_i$

$$\begin{aligned} \frac{\partial a}{\partial x_i}(x, \xi) &= \int_0^1 f_i(\xi + t(x - \xi)) dt + \int_0^1 \sum_{j=1}^n (x_j - \xi_j) \frac{\partial f_j[\xi + t(x - \xi)]}{\partial x_i} dt \\ &= \int_0^1 f_i(\xi + t(x - \xi)) dt + \int_0^1 \sum_{j=1}^n (x_j - \xi_j) \sum_{k=1}^n \frac{\partial f_j}{\partial x_k}(\xi + t(x - \xi)) \\ &\quad \cdot \frac{\partial(\xi_k + t(x_k - \xi_k))}{\partial x_i} dt \\ &= \int_0^1 f_i(\xi + t(x - \xi)) dt + \int_0^1 \sum_{j=1}^n (x_j - \xi_j) \sum_{k=1}^n t \delta_{ik} \frac{\partial f_j}{\partial x_k}(\xi + t(x - \xi)) dt \\ &= \int_0^1 f_i(\xi + t(x - \xi)) dt + \int_0^1 t \sum_{j=1}^n (x_j - \xi_j) \frac{\partial f_j}{\partial x_i}(\xi + t(x - \xi)) dt \end{aligned}$$

$$\begin{aligned} (3.8) \quad & \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a}{\partial x_i}(x, \xi) \\ &= \sum_{i=1}^n (x_i - \xi_i) \int_0^1 f_i(\xi + t(x - \xi)) dt + \int_0^1 t \sum_{j=1}^n (x_j - \xi_j) \sum_{i=1}^n (x_i - \xi_i) \frac{\partial f_j}{\partial x_i}(\xi + t(x - \xi)) dt \\ &= \sum_{i=1}^n (x_i - \xi_i) \int_0^1 f_i(\xi + t(x - \xi)) dt + \int_0^1 t \sum_{j=1}^n (x_j - \xi_j) df_j(\xi + t(x - \xi)) \\ &= \sum_{j=1}^n (x_j - \xi_j) f_j(x) \end{aligned}$$

Let  $a_0(x, \xi) = e^{a(x, \xi)}$  as in (3.3). Then

$$\begin{aligned}\frac{\partial a_0}{\partial x_i}(x, \xi) &= \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \\ \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_0}{\partial x_i}(x, \xi) &= \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \\ &= \sum_{j=1}^n (x_j - \xi_j) f_j(x) a_0(x, \xi) \quad \text{in view of (3.8)}\end{aligned}$$

So equation (3.6) is satisfied. Let

$$\begin{aligned}g_k(x, \xi) &= \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) - \sum_{i=1}^n f_i(x) \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) - \frac{1}{2} \left( \sum_{i=1}^m h_i^2(x) \right) a_{k-1}(x, \xi) \\ &\quad - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) a_{k-1}(x, \xi)\end{aligned}$$

as in (3.4) and for  $k \geq 1$

$$a_k(x, \xi) = e^{a(x, \xi)} \int_0^1 t^{k-1} e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi + t(x - \xi), \xi) dt$$

Then

$$\begin{aligned}\frac{\partial a_k}{\partial x_i}(x, \xi) &= \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \int_0^1 t^{k-1} e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi + t(x - \xi), \xi) dt \\ &\quad + e^{a(x, \xi)} \int_0^1 t^{k-1} \frac{\partial}{\partial x_i} F(\xi + t(x - \xi), \xi) dt \\ &\quad \text{where } F(\xi + t(x - \xi), \xi) = e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi + t(x - \xi), \xi) \\ &= \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \int_0^1 t^{k-1} e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi + t(x - \xi), \xi) dt \\ &\quad + e^{a(x, \xi)} \int_0^1 t^{k-1} \frac{\partial F}{\partial x_i}(\xi + t(x - \xi), \xi) \frac{\partial}{\partial x_i}(\xi_j + t(x_j - \xi_j)) dt \\ &= \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \int_0^1 t^{k-1} e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi + t(x - \xi), \xi) dt \\ &\quad + e^{a(x, \xi)} \int_0^1 t^k \frac{\partial F}{\partial x_i}(\xi + t(x - \xi), \xi) dt\end{aligned}$$



$$\begin{aligned}
& \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) \\
&= \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \int_0^1 t^{k-1} F(\xi + t(x - \xi), \xi) dt \\
&\quad + e^{a(x, \xi)} \int_0^1 t^k \sum_{i=1}^n (x_i - \xi_i) \frac{\partial F}{\partial x_i}(\xi + t(x - \xi), \xi) dt \\
&= \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_k(x, \xi) + e^{a(x, \xi)} \int_0^1 t^k dF(\xi + t(x - \xi), \xi) \\
&= \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_k(x, \xi) + e^{a(x, \xi)} F(x, \xi) \\
&\quad - k e^{a(x, \xi)} \int_0^1 t^{k-1} F(\xi + t(x - \xi), \xi) dt \\
&= \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_k(x, \xi) + e^{a(x, \xi)} e^{-a(x, \xi)} g_k(x, \xi) - k a_k(x, \xi) \\
&\Rightarrow \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) = \sum_{i=1}^n (x_i - \xi_i) f_i(x) a_k(x, \xi) + g_k(x, \xi) - k a_k(x, \xi) \\
&\Rightarrow \left( k - \sum_{i=1}^n (x_i - \xi_i) f_i(x) \right) a_k(x, \xi) + \sum_{i=1}^n (x_i - \xi_i) \frac{\partial a_k}{\partial x_i}(x, \xi) = g_k(x, \xi)
\end{aligned}$$

So equation (3.7) is also satisfied.

**Q.E.D.**

**Lemma 3** Let  $\tilde{a}_0(x, \xi) = 1$  and  $\tilde{a}_{k-1}(x, \xi) = e^{-a(x, \xi)} a_{k-1}(x, \xi)$ . Let  $\tilde{g}_k(x, \xi) = e^{-a(x, \xi)} g_k(x, \xi)$ . Then

$$\tilde{a}_k(x, \xi) = \int_0^1 t^{k-1} \tilde{g}_k(\xi + t(x - \xi), \xi) dt$$

where

$$\begin{aligned}
\tilde{g}_k(x, \xi) &= \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 \tilde{a}_{k-1}}{\partial x_i^2}(x, \xi) + \sum_{i=1}^n \left( \frac{\partial a}{\partial x_i}(x, \xi) - f_i(x) \right) \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \\
&\quad + \left[ \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a}{\partial x_i^2}(x, \xi) + \frac{1}{2} \sum_{i=1}^n \left( \frac{\partial a}{\partial x_i}(x, \xi) \right)^2 - \sum_{i=1}^n f_i(x) \frac{\partial a}{\partial x_i}(x, \xi) \right. \\
&\quad \left. - \frac{1}{2} \left( \sum_{i=1}^n h_i^2(x) \right) - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) \right] \tilde{a}_{k-1}(x, \xi)
\end{aligned}$$

**Proof:**  $a_{k-1}(x, \xi) = e^{a(x, \xi)} \tilde{a}_{k-1}(x, \xi)$

$$\begin{aligned} \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) &= e^{a(x, \xi)} \left[ \frac{\partial a}{\partial x_i}(x, \xi) \tilde{a}_{k-1}(x, \xi) + \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \right] \\ \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) &= \frac{\partial a}{\partial x_i}(x, \xi) e^{a(x, \xi)} \left[ \frac{\partial a}{\partial x_i}(x, \xi) \tilde{a}_{k-1}(x, \xi) + \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \right] \\ &\quad + e^{a(x, \xi)} \left[ \frac{\partial^2 a}{\partial x_i^2}(x, \xi) \tilde{a}_{k-1}(x, \xi) + \frac{\partial a}{\partial x_i}(x, \xi) \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) + \frac{\partial^2 \tilde{a}_{k-1}}{\partial x_i^2}(x, \xi) \right] \\ &= e^{a(x, \xi)} \left[ \left( \frac{\partial^2 a}{\partial x_i^2} + \left( \frac{\partial a}{\partial x_i}(x, \xi) \right)^2 \right) \tilde{a}_{k-1}(x, \xi) + 2 \frac{\partial a}{\partial x_i}(x, \xi) \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \right. \\ &\quad \left. + \frac{\partial^2 \tilde{a}_{k-1}}{\partial x_i^2}(x, \xi) \right] \end{aligned}$$

$$\begin{aligned} \tilde{g}_k(x, \xi) &= e^{-a(x, \xi)} g_k(x, \xi) \\ &= e^{-a(x, \xi)} \left[ \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a_{k-1}}{\partial x_i^2}(x, \xi) - \sum_{i=1}^n f_i(x) \frac{\partial a_{k-1}}{\partial x_i}(x, \xi) \right. \\ &\quad \left. - \frac{1}{2} \left( \sum_{i=1}^m h_i^2(x) \right) a_{k-1}(x, \xi) - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) a_{k-1}(x, \xi) \right] \\ &= \frac{1}{2} \sum_{i=1}^n \left[ \left( \frac{\partial^2 a}{\partial x_i^2}(x, \xi) + \left( \frac{\partial a}{\partial x_i}(x, \xi) \right)^2 \right) \tilde{a}_{k-1}(x, \xi) + 2 \frac{\partial a}{\partial x_i}(x, \xi) \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \right. \\ &\quad \left. + \frac{\partial^2 \tilde{a}_{k-1}}{\partial x_i^2}(x, \xi) \right] - \sum_{i=1}^n f_i(x) \left[ \frac{\partial a}{\partial x_i}(x, \xi) \tilde{a}_{k-1}(x, \xi) + \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \right] \\ &\quad - \frac{1}{2} \left( \sum_{i=1}^m h_i^2(x) \right) \tilde{a}_{k-1}(x, \xi) - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) \tilde{a}_{k-1}(x, \xi) \\ &= \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 \tilde{a}_{k-1}}{\partial x_i^2}(x, \xi) + \sum_{i=1}^n \left( \frac{\partial a}{\partial x_i}(x, \xi) - f_i(x) \right) \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, \xi) \\ &\quad + \left[ \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a}{\partial x_i^2}(x, \xi) + \frac{1}{2} \sum_{i=1}^n \left( \frac{\partial a}{\partial x_i}(x, \xi) \right)^2 - \sum_{i=1}^n f_i(x) \frac{\partial a}{\partial x_i}(x, \xi) - \frac{1}{2} \left( \sum_{i=1}^m h_i^2(x) \right) \right. \\ &\quad \left. - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) \right] \tilde{a}_{k-1}(x, \xi) \end{aligned}$$

On the other hand

$$\begin{aligned} a_k(x, \xi) &= e^{a(x, \xi)} \int_0^1 t^{k-1} e^{-a(\xi+t(x-\xi), \xi)} g_k(\xi+t(x-\xi), \xi) dt \\ \Rightarrow \tilde{a}_k(x, \xi) &= \int_0^1 t^{k-1} \tilde{g}_k(\xi+t(x-\xi), \xi) dt \end{aligned}$$

**Q.E.D.**

**Theorem 4** The equation (1.0) has a formal solution on  $R^n$ . In fact the solution is of the following form

$$(3.9) \quad u(t, x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{(\sqrt{2\pi})^n} t^{-n/2} \exp\left(-\frac{1}{2t} \sum_{j=1}^n (x_j - y_j)^2 + \int_0^1 \sum_{i=1}^n (x_i - y_i) f_i(y + t(x - y)) dt\right) \cdot [1 + \tilde{a}_1(x, y)t + \tilde{a}_2(x, y)t^2 + \cdots + \tilde{a}_k(x, y)t^k + \cdots] \sigma_0(y) dy_1 \cdots dy_n$$

where  $\tilde{a}_k(x, y) = \int_0^1 t^{k-1} \tilde{g}_k(y + t(x - y), y) dt$  and

$$\begin{aligned} \tilde{g}_k(x, y) = & \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 \tilde{a}_{k-1}}{\partial x_i^2}(x, y) + \sum_{i=1}^n \left( \frac{\partial a}{\partial x_i}(x, y) - f_i(x) \right) \frac{\partial \tilde{a}_{k-1}}{\partial x_i}(x, y) \\ & + \left[ \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 a}{\partial x_i^2}(x, y) + \frac{1}{2} \sum_{i=1}^n \left( \frac{\partial a}{\partial x_i}(x, y) \right)^2 - \sum_{i=1}^n f_i(x) \frac{\partial a}{\partial x_i}(x, y) \right. \\ & \left. - \frac{1}{2} \left( \sum_{i=1}^m h_i^2(x) \right) - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(x) \right] \tilde{a}_{k-1}(x, y) \end{aligned}$$

## References

- [Be] V. Benes, Exact finite dimensional filters for certain diffusions with nonlinear drift. *Stochastics*, 5 (1981), 65-92.
- [Br - Cl] R.W. Brockett and J.M.C. Clark, The geometry of the conditional density functions, in *Analysis and Optimization of Stochastic Systems*, O.L.R. Jacobs, et al, eds., Academic Press, New York, (1980), 299-309.
- [Br] R.W. Brockett, Nonlinear systems and nonlinear estimation theory, in *The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981.
- [Ch - Mi] M. Chaleyat-Maurel and D. Michel, Des resultats de non-existence de filtre de dimension finie, *Stochastics*, 13 (1984), 83-102.
- [Ch - Ya] W.L. Chiou and S.S.-T. Yau, Finite dimensional filters with nonlinear drift II: Brockett's problem on Classification of finite dimensional estimation algebras (submitted).
- [C - L - Y] J. Chen, C.W. Leung and S.S.-T. Yau, Finite dimensional filters with nonlinear drift IV: Classification of finite dimensional estimation algebras of maximal rank with state space dimension 3 (submitted for publication).
- [Co] P.C. Collingwood, Some remarks on estimation algebras, *Systems Control Lett.*, 7 (1986), 217-224.
- [Da] M.H.A. Davis, On a multiplicative functional transformation arising in nonlinear filtering theory, *Z. Wahrsch. Gebiete*, 54 (1980), 125-139.
- [Da - Ma] M.H.A. Davis and S.I. Marcus, An introduction to nonlinear filtering, in *The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981.

- [DTWY] R.T. Dong, L.F. Tam, W.S. Wong and S. S.-T. Yau, Structure and classification theorems of finite dimensional exact estimation algebras, SIAM J. Control and Optimization, Vol.29, No.4, pp.866-877, July 1991
- K - K] M. Fujisaki, G. Kallianpur and H. Kunita, Stochastic Differential Equations for the Nonlinear Filtering Problem, Osaka J. of Math., Vol.1, (1972), 19-40
- [Mi] S.K. Mitter, On the analogy between mathematical problems of nonlinear filtering and quantum physics, Ricerche di Automatica 10 (2) (1979) 163-216.
- [Oc] D.L. Ocone, Finite dimensional estimation algebras in nonlinear filtering, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981..
- [St] S. Steinberg, Applications of the Lie algebraic formulas of Baker, Campbell, Hausdorff and Zassenhaus to the calculation of explicit solutions of partial differential equations, J. Differential Equations, 26 (1979), 404-434.
- W - Y] L.F. Tam, W.S. Wong and S. S.-T. Yau, On a necessary and sufficient condition for finite dimensionality of estimation algebras, SIAM J. Control and Optimization, Vol. 28, No. 1 (1990), 173-185.
- We - No] J. Wei and E. Norman, On global representations of the solutions of linear differential equations as a product of exponentials, Proc. Amer. Math. Soc., 15 (1964), 327-334.
- [Wi] D.V. Widder, The Heat Equation, Mathematics 67, Academy Press, 1975
- [Wo1] W.S. Wong, New classes of finite dimensional nonlinear filters, Systems Control Lett., 3 (1983), 155-164.
- [Wo2] W.S. Wong, On a new class of finite dimensional estimation algebras, Systems Control Lett., 9 (1987), 79-83.
- [Wo3] W.S. Wong, Theorems on the structure of finite dimensional estimation algebras, Systems Control Lett., 9 (1987), 117-124.
- Ya - Ch] S. S.-T. Yau and W.L. Chiou, Recent results on classification of finite dimensional estimation algebras : Dimension of State Space  $\leq 2$ , Proceedings of the 30th IEEE Conference on Decision and Control, Brighton, England, Dec 11-13, 1991
- [Ya1] S. S.-T. Yau, Recent results on nonlinear filtering, New class of finite dimensional filters, Proceedings of the 29th IEEE Conference on Decision and Control, Honolulu, Hawaii, Dec.5-7, (1990) 231-233
- [Ya2] S. S.-T. Yau, Finite dimensional filters with nonlinear drift I : A class of filters including both Kalman-Bucy filters and Benes filters (to appear) *J. of Math. Systems, Estimation and Control*

# NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS OF INTEREST IN NONLINEAR OPTICS\*

M. J. POTASEK  
Department of Applied Physics  
Columbia University  
New York, NY 10027

**ABSTRACT.** The analysis of nonlinear optical phenomena is important because of the wide range of potential applications in various fields of science and engineering. Photonics is becoming increasingly significant and research in this area is of vital importance for applications ranging from optical computing to novel light sources. It has been recognized that solitons may be natural bits of information for gating and switching. While considerable effort has focussed on the picosecond time domain, research is now evolving in the femtosecond time domain. We have calculated the performance of the first femtosecond soliton all-optical switch.

**INTRODUCTION.** Because of their rapid response time, nonlinear optical materials are gathering considerable interest. Much of this research has involved picosecond or greater duration pulses. However, the recent development of femtosecond light sources in the visible and near infrared spectral regions makes possible the exploration of new phenomena on ultrashort time scales. Research into the fundamental aspects of ultrashort pulses is of interest for possible femtosecond soliton lasers, amplifiers, and optical computers. Femtosecond pulses are now available from a number of optical lasers, such as the NaCl color center laser, mode-locked Er fiber laser and the colliding pulse mode-locked semiconductor laser. However, mathematical analysis is now just evolving for this rapidly developing field and has not kept pace with experiments. The emphasis on shorter pulses requires extension beyond traditional techniques. Light propagating in a dielectric medium is given by

$$\nabla^2 E - \frac{1}{\epsilon_0 c^2} \frac{\partial^2 D}{\partial t^2} - \nabla(\nabla \cdot E) = 0 \quad (1)$$

\*Supported in part by the U.S. Army Research Office.

where  $E$  is the electromagnetic field and  $D$  is the displacement vector which is expanded in a power series in  $E$ ,

$$D = \epsilon_0 \int_{-\infty}^t dt_1 \chi_1(t-t_1) E(t_1) \quad (2)$$

$$+ \epsilon_0 \int_{-\infty}^t dt_1 \int_{-\infty}^t dt_2 \int_{-\infty}^t dt_3 \chi_3(t-t_1, t-t_2, t-t_3) E(t_1) E(t_2) E(t_3)$$

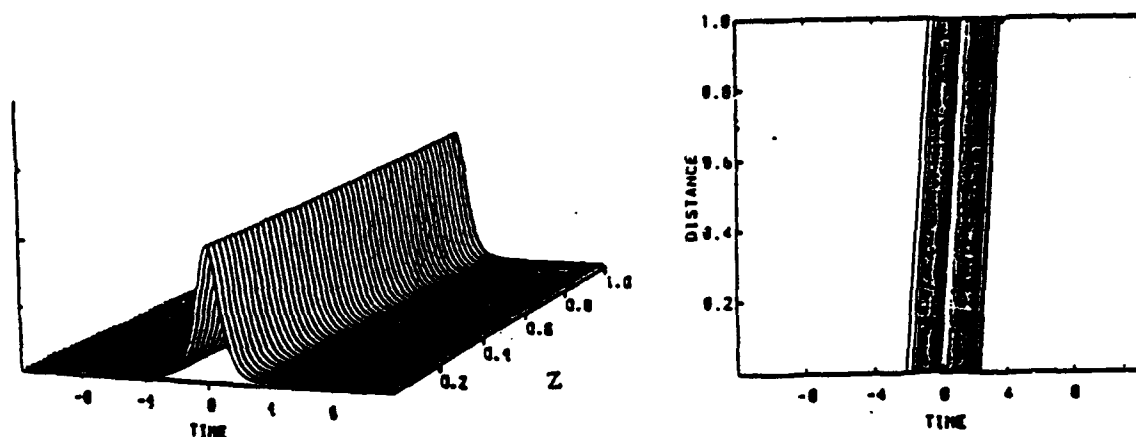
For femtosecond pulses, one obtains

$$iq_z + \frac{1}{2} q_{zz} + |q|^2 q + i\epsilon q_{zz} + i\epsilon_2 |q|^2 q_t + i\epsilon_3 q^2 q_t^* - \epsilon_4 |q|^2 q_t = 0 \quad (3)$$

where  $q$  is the dimensionless slowly varying envelope of the electromagnetic field, the subscripts  $z$  and  $t$  refer to differentiation with respect to space and time, respectively, and the coefficients depends upon various dispersive and nonlinear constants. For certain physical conditions, Eq. (3) reduces to

$$iq_z + \frac{1}{2} q_{zz} + |q|^2 q + i\epsilon (q_{zz} + 6|q|^2 q_t) = 0 \quad (4)$$

The figure below shows the propagation of the soliton.



**FEMTOSECOND ALL-OPTICAL SWITCHES.** An important area for optical computing and communications is intensity dependent optical switching.

Optical switching is dependent on the nonlinear index of refraction of optical materials. Device configuration consists of nonlinear directional couplers or birefringent couplers. Experimentally all-optical logic devices including an inverter, exclusive-OR and AND logic gates were demonstrated in the picosecond time domain. This presentation describes the mathematical results for novel femtosecond optical switches.

The coupled equations are given by

$$iq_{1z} + \frac{1}{2}q_{1z} + (|q_1|^2 + |q_2|^2)q_1 + kq_2 + i\epsilon \left[ q_{1zz} + 3(|q_1|^2 + |q_2|^2)q_{1z} + 3(q_1^*q_{1z} + q_2^*q_{2z})q_1 \right] = 0 \quad (5)$$

$$iq_{2z} + \frac{1}{2}q_{2z} + (|q_1|^2 + |q_2|^2)q_2 + kq_1 + i\epsilon \left[ q_{2zz} + 3(|q_1|^2 + |q_2|^2)q_{2z} + 3(q_1^*q_{1z} + q_2^*q_{2z})q_2 \right] = 0 \quad (6)$$

The equations can be solved by the inverse scattering transform method. For the  $N=1$  soliton one obtains

$$q_1 = \frac{1}{\sqrt{2}} \left( \sin\theta e^{i(kz - \phi_1)} - \cos\theta e^{-i(kz + \phi_2)} \right) q(z,t) \quad (7)$$

$$q_2 = \frac{1}{\sqrt{2}} \left( \sin\theta e^{i(kz - \phi_1)} + \cos\theta e^{-i(kz + \phi_2)} \right) q(z,t) \quad (8)$$

where

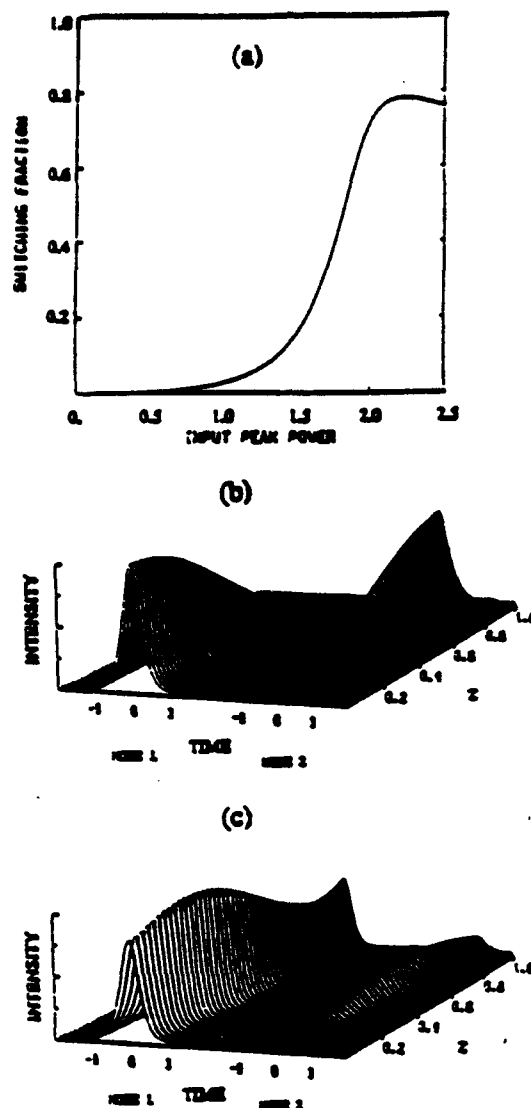
$$q(z,t) = 2\eta \operatorname{sech} [2\eta(t + (4\epsilon(\eta^2 - 3\xi^2) + 2\xi)z)] \exp [-i(2\xi t - 8\epsilon\xi(\xi^2 - 3\eta^2) + 2(\xi^2 - \eta^2)z)] ,$$

$$p = \xi + i\eta .$$

The angles describe the relative intensity and phase of the input waves.

A figure of the optical switching behavior is shown below. Part (a) shows the output switching fraction as a function of input power, part (b) shows the optical switching behavior at low input intensity

and part (c) shows the optical switching behavior at high input intensity.



**CONCLUSIONS.** These results have been correlated with experimental parameters. Experiments are planned to verify the concepts. This would be the first demonstration of femtosecond soliton all-optical switching which could advance the frontier of optical devices by several orders of magnitude and constitute a breakthrough of the physical understanding of high data rate systems..

#### Acknowledgement

I thank J. M. Fang for numerical programs and R. Tascal for inverse scattering results.



# Scattering Algorithms for Beltrami Fields

Balasubramaniam Shanker and Akhlesh Lakhtakia

*Department of Engineering Science and Mechanics*

*227, Hammond Building*

*Pennsylvania State University*

*University Park, PA 16802-1401*

## **Abstract:**

*Beltrami fields yield a convenient representation of the electromagnetic fields in a chiral medium, and the scattering of electromagnetic waves by a scatterer immersed in a chiral host medium, is easily tackled using Beltrami fields. In this work, volume integral equations are set up using equivalent Beltrami current sources to represent a chiral scatterer in a chiral medium. These equations are then converted into algebraic equations and the scattering algorithms obtained, using both the method of moments (MOM) and the coupled dipole method (CDM). The singular behavior of the dyadic Green's function for Beltrami fields in the neighborhood of the source point is estimated with care, and the strong and the weak forms of both the MOM and the CDM algorithms are derived.*

## **1. Introduction:**

A Beltrami field is defined as one that is proportional to its curl in a source-free region. Nature provides us with typical examples; hurricanes, water spouts and vortex flows can be very well approximated by a Beltrami field. In electromagnetic theory, Beltrami fields are found as toroidal and poloidal fields and circularly polarized plane waves. Beltrami fields came to prominence in electromagnetics in 1974 due to Bohren [1], although they had been used as early as 1907 [2]. The usefulness and application of Beltrami fields cannot be doubted, as is evident from [3-5].

In the sequel we consider scattering by a chiral inclusion embedded in a chiral host medium. The electromagnetic fields that are incident on the inclusion are expressed in terms of Beltrami fields, and we derive expressions for the scattered Beltrami fields. But before we formulate the problem, the mathematical concept of Beltrami fields is introduced in the next section. In section 3, the electric and magnetic fields and the volume current densities are expressed in terms of their Beltrami equivalents. The scatterer is then replaced by equivalent Beltrami current densities and volume integral equations are obtained for scattering. Finally, in section 4 these integral equations are reduced to algebraic ones and two algorithms, the method of moments and the coupled dipole method, are developed.

## 2. Beltrami fields

A Beltrami field  $\mathbf{Q}(\mathbf{r})$  radiated by a Beltrami source density  $\mathbf{W}(\mathbf{r})$  satisfies the relation

$$\nabla \times \mathbf{Q}(\mathbf{r}) + \kappa \mathbf{Q}(\mathbf{r}) = \mathbf{W}(\mathbf{r}), \quad (1)$$

where the real and imaginary part of  $\kappa$  are of the same sign for physical validity.

The solution to such an equation, for any  $\mathbf{r}$ , is

$$\mathbf{Q}(\mathbf{r}) = \mathbf{Q}_h(\mathbf{r}) + \iiint_V \underline{\underline{G}}(\mathbf{r}, \mathbf{r}') \bullet \mathbf{W}(\mathbf{r}') d^3\mathbf{r}', \quad (2)$$

where  $\mathbf{Q}_h(\mathbf{r})$  is the solution to  $\nabla \times \mathbf{Q}(\mathbf{r}) + \kappa \mathbf{Q}(\mathbf{r}) = 0$  everywhere. The Green's dyadic  $\underline{\underline{G}}(\mathbf{r}, \mathbf{r}')$  for the Beltrami field satisfies the equation

$$\nabla \times \underline{\underline{G}}(\mathbf{r}, \mathbf{r}') + \kappa \underline{\underline{G}}(\mathbf{r}, \mathbf{r}') = \mathbf{I} \delta(\mathbf{R}), \quad \mathbf{R} = \mathbf{r} - \mathbf{r}', \quad (3)$$

where  $\delta(\mathbf{R})$  is the Dirac delta and  $\mathbf{I}$  is the identity dyadic. The Green's function has been obtained by a number of approaches [6-8] and turns out to be

$$\underline{\underline{G}}(\mathbf{r}, \mathbf{r}') = -(\kappa \mathbf{I} + \nabla \nabla / \kappa - \nabla \times \mathbf{I}) g(\mathbf{r}, \mathbf{r}'), \quad (4)$$

where

$$g(\mathbf{r}, \mathbf{r}') = \exp(\pm i\kappa|\mathbf{R}|)/4\pi|\mathbf{R}|, \quad (5)$$

is the scalar Green's function commonly used in electromagnetics and acoustics. In (5), the upper sign is to be used for  $\text{Re}\{\kappa\} > 0$  and the lower for  $\text{Re}\{\kappa\} < 0$ .

Due to the presence of the  $\nabla\nabla$  term in (4) the integrand in (2) is singular at  $\mathbf{R} = \mathbf{r} - \mathbf{r}'$ . But for an electrically small region  $V$ , whose cross-sectional dimensions are much smaller than the wavelength  $2\pi/|\kappa|$ , the field at the source point can be approximated as [9]

$$\mathbf{Q}(\mathbf{r}') = \mathbf{Q}_h(\mathbf{r}') - \{\kappa \underline{\underline{M}} - \underline{\underline{L}}/\kappa\} \cdot \mathbf{W}(\mathbf{r}'), \quad \mathbf{r} \in V, \quad (6a)$$

where

$$\underline{\underline{L}} = (1/4\pi) \iint_{\partial V} [\mathbf{n}' \mathbf{R}/|\mathbf{R}|^3] d^2 \mathbf{r}', \quad (6b)$$

$$\begin{aligned} \underline{\underline{M}} = (1/4\pi) \iiint_V \left( [\underline{\underline{I}} - \nabla \times \underline{\underline{I}}/\kappa] \{e^{i\kappa|\mathbf{R}|}/|\mathbf{R}|\} \right. \\ \left. + \nabla \nabla [(e^{i\kappa|\mathbf{R}|} - 1)/\kappa^2 |\mathbf{R}|] \right) d^3 \mathbf{r}', \end{aligned} \quad (6c)$$

with  $\mathbf{n}'$  being the unit outward normal to the surface  $\partial V$  of  $V$  at  $\mathbf{r}' \in \partial V$ .

### 3. Field representation

#### 3.1 Chiral host medium

With these preliminaries, we now look at fields in a chiral medium. In the Drude-Born-Federov representation, the frequency domain constitutive relations for a chiral medium are

$$\mathbf{D} = \epsilon_a \mathbf{E} + \epsilon_a \beta_a \nabla \times \mathbf{E}, \quad \mathbf{B} = \mu_a \mathbf{H} + \mu_a \beta_a \nabla \times \mathbf{H}, \quad (7a,b)$$

where  $\epsilon_a$  and  $\mu_a$  are the permittivity and the permeability scalars, respectively, and  $\beta_a$  is chirality parameter of the host medium. An  $\exp[-i\omega t]$  time is implicitly assumed throughout this analysis. On transforming the  $\mathbf{E}$  and  $\mathbf{H}$  fields into Beltrami fields [10], the Maxwell curl postulates can be written as

$$\nabla \times \mathbf{Q}_1 - \gamma_1 \mathbf{Q}_1 = \mathbf{W}_1, \quad \nabla \times \mathbf{Q}_2 + \gamma_2 \mathbf{Q}_2 = \mathbf{W}_2, \quad (8a,b)$$

with the Beltrami fields given by

$$\mathbf{Q}_1 = [1/2][\mathbf{E} + i\eta_a \mathbf{H}], \quad (9a)$$

$$\mathbf{Q}_2 = [1/2][\mathbf{H} + i\mathbf{E}/\eta_a], \quad (9b)$$

and the Beltrami source current densities by

$$W_1 = [\gamma_1/2k_a] [i\eta_a J - K], \quad (9c)$$

$$W_2 = [\gamma_2/2k_a] [-i(1/\eta_a) K + J]. \quad (9d)$$

The two wavenumbers in the host medium are found to be

$$\gamma_1 = k_a/(1-k_a\beta_a), \quad (10a)$$

$$\gamma_2 = k_a/(1+k_a\beta_a), \quad (10b)$$

where  $k_a = \omega\sqrt{\epsilon_a\mu_a}$  is used for convenience in notation; and  $\eta_a = \sqrt{\mu_a/\epsilon_a}$  is the impedance of the chiral medium. It is assumed here that  $\text{Re}\{\gamma_1, \gamma_2\} > 0$ ,  $\text{Im}\{\gamma_1, \gamma_2\} \geq 0$  and  $\text{Re}\{\eta_a\} > 0$ ; and it is also noted that  $Q_1$  is a left-handed field and  $Q_2$  is a right-handed field [1, 11].

Thus, to solve for the Beltrami fields  $Q_1(r)$  and  $Q_2(r)$ , suitable modifications of (2) yield

$$Q_1(r) = Q_{1h}(r) + (\gamma_1 + \gamma_2) \iiint_V \underline{\underline{G}}_1(r, r') \bullet W_1(r') d^3r', \quad (11a)$$

$$Q_2(r) = Q_{2h}(r) - (\gamma_1 + \gamma_2) \iiint_V \underline{\underline{G}}_2(r, r') \bullet W_2(r') d^3r', \quad (11b)$$

where  $V$  now is a region not necessarily small in electrical size, while the dyadics

$$\begin{aligned} \underline{\underline{G}}_1(r, r') &= (\gamma_1 + \gamma_2)^{-1} \underline{\underline{G}}(r, r')|_{\kappa=-\gamma_1} \\ &= (\gamma_1 \underline{\underline{I}} + \nabla \nabla / \gamma_1 + \nabla \times \underline{\underline{I}}) \{ \exp(i\gamma_1 |R|) / 4\pi |R| \} / (\gamma_1 + \gamma_2), \end{aligned} \quad (12a)$$

$$\begin{aligned} \underline{\underline{G}}_2(r, r') &= -(\gamma_1 + \gamma_2)^{-1} \underline{\underline{G}}(r, r')|_{\kappa=\gamma_2} \\ &= (\gamma_2 \underline{\underline{I}} + \nabla \nabla / \gamma_2 - \nabla \times \underline{\underline{I}}) \{ \exp(i\gamma_2 |R|) / 4\pi |R| \} / (\gamma_1 + \gamma_2). \end{aligned} \quad (12b)$$

### 3.2 Equivalent current densities

Equations (11a,b) are volume integral equations which can be used to solve for radiation from a given source distribution. If, however, we consider scattering,  $Q_{1h}(r)$  and  $Q_{2h}(r)$  are interpreted as the fields that exist in the absence of the scatterer and are, therefore, the incident fields. Our task now is to replace the scatterer by current densities which are equivalent to it.

Let all space be divided into two mutually exclusive regions  $V_{\text{ext}}$  and  $V_{\text{int}}$  which represent the host and the inclusion media, respectively. The region  $V_{\text{int}}$  is filled with chiral matter obeying the constitutive relations

$$\mathbf{D}(\mathbf{r}) = \epsilon_b \mathbf{E}(\mathbf{r}) + \epsilon_b \beta_b \nabla \times \mathbf{E}(\mathbf{r}), \quad \mathbf{r} \in V_{\text{int}}, \quad (13a)$$

$$\mathbf{B}(\mathbf{r}) = \mu_b \mathbf{H}(\mathbf{r}) + \mu_b \beta_b \nabla \times \mathbf{H}(\mathbf{r}), \quad \mathbf{r} \in V_{\text{int}}, \quad (13b)$$

where  $\epsilon_b$  and  $\mu_b$  are the permittivity and the permeability, respectively, and  $\beta_b$  is the chirality parameter of the inclusion material.

The source-free Maxwell postulates can be written everywhere as

$$\begin{aligned} \nabla \times \mathbf{E}(\mathbf{r}) - [k_a^2 \beta_a \mathbf{E}(\mathbf{r}) + i\omega \mu_a \mathbf{H}(\mathbf{r})] / (1 - k_a^2 \beta_a^2) \\ = [-\mathbf{K}_{\text{eq}}(\mathbf{r}) + i\omega \mu_a \beta_a \mathbf{J}_{\text{eq}}(\mathbf{r})] / (1 - k_a^2 \beta_a^2), \quad \mathbf{r} \in V_{\text{ext}} + V_{\text{int}}, \end{aligned} \quad (14a)$$

$$\begin{aligned} \nabla \times \mathbf{H}(\mathbf{r}) - [k_a^2 \beta_a \mathbf{H}(\mathbf{r}) - i\omega \epsilon_a \mathbf{E}(\mathbf{r})] / (1 - k_a^2 \beta_a^2) \\ = [\mathbf{J}_{\text{eq}}(\mathbf{r}) + i\omega \epsilon_a \beta_a \mathbf{K}_{\text{eq}}(\mathbf{r})] / (1 - k_a^2 \beta_a^2), \quad \mathbf{r} \in V_{\text{ext}} + V_{\text{int}}. \end{aligned} \quad (14b)$$

Here,

$$\mathbf{J}_{\text{eq}}(\mathbf{r}) = 0, \quad \mathbf{r} \in V_{\text{ext}}, \quad (15a)$$

$$\mathbf{K}_{\text{eq}}(\mathbf{r}) = 0, \quad \mathbf{r} \in V_{\text{ext}}, \quad (15b)$$

$$\mathbf{J}_{\text{eq}}(\mathbf{r}) = i\omega [a_{\text{ee}} \mathbf{E}(\mathbf{r}) + a_{\text{eh}} \mathbf{H}(\mathbf{r})], \quad \mathbf{r} \in V_{\text{int}}, \quad (15c)$$

$$\mathbf{K}_{\text{eq}}(\mathbf{r}) = i\omega [a_{\text{he}} \mathbf{E}(\mathbf{r}) + a_{\text{hh}} \mathbf{H}(\mathbf{r})], \quad \mathbf{r} \in V_{\text{int}}, \quad (15d)$$

and the quantities

$$a_{\text{ee}} = \epsilon_a \beta_a \left[ \frac{k_b^2 \beta_b}{1 - k_b^2 \beta_b^2} - \frac{k_a^2 \beta_a}{1 - k_a^2 \beta_a^2} \right] + \left[ \frac{\epsilon_a}{1 - k_a^2 \beta_a^2} - \frac{\epsilon_b}{1 - k_b^2 \beta_b^2} \right], \quad (16a)$$

$$a_{\text{eh}} = i\omega \epsilon_a \beta_a \left[ \frac{\mu_b}{1 - k_b^2 \beta_b^2} - \frac{\mu_a}{1 - k_a^2 \beta_a^2} \right] + \frac{1}{i\omega} \left[ \frac{k_b^2 \beta_b}{1 - k_b^2 \beta_b^2} - \frac{k_a^2 \beta_a}{1 - k_a^2 \beta_a^2} \right], \quad (16b)$$

$$a_{\text{he}} = i\omega \mu_a \beta_a \left[ \frac{\epsilon_a}{1 - k_a^2 \beta_a^2} - \frac{\epsilon_b}{1 - k_b^2 \beta_b^2} \right] - \frac{1}{i\omega} \left[ \frac{k_b^2 \beta_b}{1 - k_b^2 \beta_b^2} - \frac{k_a^2 \beta_a}{1 - k_a^2 \beta_a^2} \right], \quad (16c)$$

$$a_{\text{hh}} = \mu_a \beta_a \left[ \frac{k_b^2 \beta_b}{1 - k_b^2 \beta_b^2} - \frac{k_a^2 \beta_a}{1 - k_a^2 \beta_a^2} \right] - \left[ \frac{\mu_b}{1 - k_b^2 \beta_b^2} - \frac{\mu_a}{1 - k_a^2 \beta_a^2} \right]. \quad (16d)$$

Knowing the electric and magnetic current densities, and using (9c,d) we can set down

$$\nabla \times \mathbf{Q}_1(\mathbf{r}) - \gamma_1 \mathbf{Q}_1(\mathbf{r}) = \mathbf{W}_{1eq}(\mathbf{r}), \mathbf{r} \in V_{ext} + V_{int}, \quad (17a)$$

$$\nabla \times \mathbf{Q}_2(\mathbf{r}) + \gamma_2 \mathbf{Q}_2(\mathbf{r}) = \mathbf{W}_{2eq}(\mathbf{r}), \mathbf{r} \in V_{ext} + V_{int}, \quad (17b)$$

where the equivalent Beltrami source current densities are specified by

$$\mathbf{W}_{1eq}(\mathbf{r}) = \mathbf{0}, \mathbf{r} \in V_{ext}, \quad (18a)$$

$$\mathbf{W}_{2eq}(\mathbf{r}) = \mathbf{0}, \mathbf{r} \in V_{ext}, \quad (18b)$$

$$\mathbf{W}_{1eq}(\mathbf{r}) = [\gamma_1/2k_a] [a_{11}\mathbf{Q}_1(\mathbf{r}) + a_{12}\mathbf{Q}_2(\mathbf{r})], \mathbf{r} \in V_{int}, \quad (18c)$$

$$\mathbf{W}_{2eq}(\mathbf{r}) = [\gamma_2/2k_a] [a_{21}\mathbf{Q}_1(\mathbf{r}) + a_{22}\mathbf{Q}_2(\mathbf{r})], \mathbf{r} \in V_{int}, \quad (18d)$$

with

$$a_{11} = i\omega \{ (a_{eh} - a_{he}) + i(\eta_a^{-1} a_{hh} + \eta_a a_{ee}) \}, \quad (19a)$$

$$a_{12} = i\omega \{ (\eta_a^2 a_{ee} - a_{hh}) + i\eta_a (a_{eh} + a_{he}) \}, \quad (19b)$$

$$a_{21} = i\omega \{ (a_{ee} - a_{hh} / \eta_a^2) - (i/\eta_a) (a_{eh} + a_{he}) \}, \quad (19c)$$

$$a_{22} = i\omega \{ (a_{eh} - a_{he}) - i(\eta_a^{-1} a_{hh} + \eta_a a_{ee}) \}. \quad (19d)$$

Thus, through (18a-d) the Beltrami field scattering problem has been altered to an equivalent Beltrami field radiation problem. As the equivalent current densities are identically null in  $V_{ext}$ , the volume integral equations for Beltrami fields take the form

$$\begin{aligned} \mathbf{Q}_1(\mathbf{r}) = \mathbf{Q}_{1inc}(\mathbf{r}) + \iiint_{V_{int}} [\underline{\underline{A}}_{11}(\mathbf{r}, \mathbf{r}') \bullet \mathbf{Q}_1(\mathbf{r}')] d^3\mathbf{r}' \\ + \iiint_{V_{int}} [\underline{\underline{A}}_{12}(\mathbf{r}, \mathbf{r}') \bullet \mathbf{Q}_2(\mathbf{r}')] d^3\mathbf{r}', \mathbf{r} \in V_{ext} + V_{int}, \end{aligned} \quad (20a)$$

$$\begin{aligned} \mathbf{Q}_2(\mathbf{r}) = \mathbf{Q}_{2inc}(\mathbf{r}) + \iiint_{V_{int}} [\underline{\underline{A}}_{21}(\mathbf{r}, \mathbf{r}') \bullet \mathbf{Q}_1(\mathbf{r}')] d^3\mathbf{r}' \\ + \iiint_{V_{int}} [\underline{\underline{A}}_{22}(\mathbf{r}, \mathbf{r}') \bullet \mathbf{Q}_2(\mathbf{r}')] d^3\mathbf{r}', \mathbf{r} \in V_{ext} + V_{int}, \end{aligned} \quad (20b)$$

with the dyadic kernels

$$\underline{\underline{A}}_{11}(\mathbf{r}, \mathbf{r}') = \frac{\gamma_1^2 \gamma_2}{k_a^2} a_{11} \underline{\underline{G}}_1(\mathbf{r}, \mathbf{r}'), \quad (21a)$$

$$\underline{\underline{A}}_{12}(\mathbf{r}, \mathbf{r}') = \frac{\gamma_1^2 \gamma_2}{k_a^2} a_{12} \underline{\underline{G}}_1(\mathbf{r}, \mathbf{r}'), \quad (21b)$$

$$\underline{\underline{A}}_{21}(\mathbf{r}, \mathbf{r}') = - \frac{\gamma_2^2 \gamma_1}{k_a^2} a_{21} \underline{\underline{G}}_2(\mathbf{r}, \mathbf{r}'), \quad (21c)$$

$$\underline{\underline{A}}_{22}(\mathbf{r}, \mathbf{r}') = - \frac{\gamma_2^2 \gamma_1}{k_a^2} \underline{\underline{a}}_{22} \underline{\underline{G}}_2(\mathbf{r}, \mathbf{r}'). \quad (21d)$$

For prescribed incident fields,  $\underline{\underline{Q}}_{1inc}$  and  $\underline{\underline{Q}}_{2inc}$ , (20a,b) have to be solved for the Beltrami fields. The results have to then be put back to find the scattered fields. To solve these volume integral equations, we use the longwavelength approximation to rewrite these equations in algebraic form and develop two scattering algorithms – the method of moments and the coupled dipole method.

#### 4.1 Method of Moments (MOM)

In 1968 Harrington [12] introduced into electromagnetic engineering community the term “method of moments” to denote an approach by which a linear operator equation is transformed into a set of algebraic equations. Since then more sophisticated versions have been developed [13,14], but the original and the simplest version suffices for our purposes.

In order to convert the volume integral equations into algebraic ones, the region  $V_{int}$  is partitioned into simply connected, non-overlapping subregions  $V_m$  ( $m = 1, 2, \dots, M$ ), each bounded by a surface  $\partial V_m$  on which a unique unit outward normal can be unambiguously prescribed, and is, therefore, at least once-differentiable. The surface of the subregion is assumed to be convex, and the maximum chord of the subregion  $V_m$  is assumed small compared to the wavelength in both  $V_m$  and  $V_{ext}$  such that longwavelength approximations hold. As a consequence of the longwavelength approximation, we can assume  $\underline{\underline{Q}}_1(\mathbf{r}) \equiv \underline{\underline{Q}}_1(\mathbf{r}_m)$  and  $\underline{\underline{Q}}_2(\mathbf{r}) \equiv \underline{\underline{Q}}_2(\mathbf{r}_m)$  for all  $\mathbf{r} \in V_m$ , where  $\mathbf{r}_m$  is a distinguished point, generally the centroid of  $V_m$ . Next, the approximate evaluations

$$\iiint_{V_n} [\underline{\underline{G}}_1(\mathbf{r}_m, \mathbf{r}') \bullet \Phi(\mathbf{r}')] d^3\mathbf{r}' \equiv v_n \underline{\underline{G}}_1(\mathbf{r}_m, \mathbf{r}_n) \bullet \Phi(\mathbf{r}_n), \quad n \neq m, \quad (22a)$$

$$\iiint_{V_m} [\underline{\underline{G}}_1(\mathbf{r}_m, \mathbf{r}') \bullet \Phi(\mathbf{r}')] d^3\mathbf{r}' \equiv \frac{k_a}{2\gamma_1^2 \gamma_2} [-\underline{\underline{L}}_m + \gamma_1^2 \underline{\underline{M}}_{1m}] \bullet \Phi(\mathbf{r}_m), \quad (22b)$$

$$\iiint_{V_n} [G_2(\mathbf{r}_m, \mathbf{r}') \bullet \Phi(\mathbf{r}')] d^3\mathbf{r}' \equiv v_n G_2(\mathbf{r}_m, \mathbf{r}_n) \bullet \Phi(\mathbf{r}_n), n \neq m, \quad (22c)$$

$$\iiint_{V_m} [G_2(\mathbf{r}_m, \mathbf{r}') \bullet \Phi(\mathbf{r}')] d^3\mathbf{r}' \equiv \frac{k_a}{2\gamma_2^2 \gamma_1} [-\underline{\underline{L}}_m + \gamma_2^2 \underline{\underline{M}}_{2m}] \bullet \Phi(\mathbf{r}_m), \quad (22d)$$

follow, with  $v_n$  being the volume of  $V_n$ , and  $\Phi$  representing either  $Q_1$ , or  $Q_2$  in the subregion of integration and the dyadics

$$\underline{\underline{L}}_m = (1/4\pi) \iint_{\partial V_m} [n' \mathbf{R}_m / |\mathbf{R}_m|^3] d^2\mathbf{r}', \quad (23a)$$

$$\begin{aligned} \underline{\underline{M}}_{1m} = (1/4\pi) \iiint_{V_m} \left( [\underline{\underline{I}} + \nabla \times \underline{\underline{I}} / \gamma_1] \{e^{i\gamma_1 |\mathbf{R}_m|} / |\mathbf{R}_m|\} \right. \\ \left. + \nabla \nabla \{ (e^{i\gamma_1 |\mathbf{R}_m|} - 1) / \gamma_1^2 |\mathbf{R}_m| \} \right) d^3\mathbf{r}', \end{aligned} \quad (23b)$$

$$\begin{aligned} \underline{\underline{M}}_{2m} = (1/4\pi) \iiint_{V_m} \left( [\underline{\underline{I}} - \nabla \times \underline{\underline{I}} / \gamma_2] \{e^{i\gamma_2 |\mathbf{R}_m|} / |\mathbf{R}_m|\} \right. \\ \left. + \nabla \nabla \{ (e^{i\gamma_2 |\mathbf{R}_m|} - 1) / \gamma_2^2 |\mathbf{R}_m| \} \right) d^3\mathbf{r}', \end{aligned} \quad (23c)$$

with  $\mathbf{R}_m = \mathbf{r}_m - \mathbf{r}'$  and  $n'$  being the unit outward normal to the surface  $\partial V_m$  at  $\mathbf{r}' \in \partial V_m$ . We observe that the union of  $V_m$  is only approximately congruent with  $V_{int}$  in practical situations. Also to be noted is that the use of only the L dyadic constitutes the "weak" form of MOM [9], while the use of both the L and the M dyadics is the "strong" form thereof.

With these approximations, the volume integral equations can be converted into 6M algebraic equations in terms of the cartesian components of the fields  $Q_1(\mathbf{r}_n)$  and  $Q_2(\mathbf{r}_n)$ . Thus,

$$Q_{1inc}(\mathbf{r}_m) = \sum_{n=1,2,\dots,M} [\underline{\underline{B}}_{11,mn} \bullet Q_1(\mathbf{r}_n) + \underline{\underline{B}}_{12,mn} \bullet Q_2(\mathbf{r}_n)], 1 \leq m \leq M, \quad (24a)$$

$$Q_{2inc}(\mathbf{r}_m) = \sum_{n=1,2,\dots,M} [\underline{\underline{B}}_{21,mn} \bullet Q_1(\mathbf{r}_n) + \underline{\underline{B}}_{22,mn} \bullet Q_2(\mathbf{r}_n)], 1 \leq m \leq M, \quad (24b)$$

where

$$\underline{\underline{B}}_{qs,mn} = -v_n \underline{\underline{A}}_{qs}(\mathbf{r}_m, \mathbf{r}_n); q = 1, 2; s = 1, 2; m \neq n, \quad (25a)$$

$$\underline{\underline{B}}_{11,mn} = \underline{\underline{I}} - [1/2k_a] [-\underline{\underline{L}}_m + \gamma_1^2 \underline{\underline{M}}_{1m}] a_{11}, \quad (25b)$$

$$\underline{\underline{B}}_{12,mn} = -[1/2k_a] [-\underline{\underline{L}}_m + \gamma_1^2 \underline{\underline{M}}_{1m}] a_{12}, \quad (25c)$$

$$\underline{\underline{B}}_{21,mn} = [1/2k_a] [-\underline{\underline{L}}_m + \gamma_2^2 \underline{\underline{M}}_{2m}] a_{21}, \quad (25d)$$

$$\underline{\underline{B}}_{22,mn} = \underline{\underline{I}} + [1/2k_a] [-\underline{\underline{L}}_m + \gamma_2^2 \underline{\underline{M}}_{2m}] a_{22}. \quad (25e)$$



The matrix equations (24a,b) can be solved using standard matrix methods provided they are not algorithmically singular. The scattered Beltrami fields can be computed through (20a,b) as

$$\begin{aligned} Q_{1sca}(r) &= Q_1(r) - Q_{1inc}(r) \\ &= \sum_{n=1,2,\dots,M} v_n [A_{11}(r, r_n) \bullet Q_1(r_n) + A_{12}(r, r_n) \bullet Q_2(r_n)], \quad r \in V_{ext}, \end{aligned} \quad (26a)$$

$$\begin{aligned} Q_{2sca}(r) &= Q_2(r) - Q_{2inc}(r) \\ &= \sum_{n=1,2,\dots,M} v_n [A_{21}(r, r_n) \bullet Q_1(r_n) + A_{22}(r, r_n) \bullet Q_2(r_n)], \quad r \in V_{ext}. \end{aligned} \quad (26b)$$

## 4.2 Rotability dyadics

In the MOM the unknowns are the fields that are actually present at the subregion. For the development of the coupled dipole method, we need the fields that excite the  $m^{th}$  subregion. The field exciting  $V_m$  is the sum of the field that is incident on the subregion and the fields that are scattered by the other subregions. In other words, the field exciting  $V_m$  is the actual field in the subregion if it were to be occupied by the host medium. Thus, by definition

$$\begin{aligned} Q_{1exc}(r_m) &= Q_{1inc}(r_m) - \sum_{n=1,2,\dots,M; n \neq m} [B_{11,mn} \bullet Q_1(r_n) \\ &\quad + B_{12,mn} \bullet Q_2(r_n)], \quad 1 \leq m \leq M, \end{aligned} \quad (27a)$$

$$\begin{aligned} Q_{2exc}(r_m) &= Q_{2inc}(r_m) - \sum_{n=1,2,\dots,M; n \neq m} [B_{21,mn} \bullet Q_1(r_n) \\ &\quad + B_{22,mn} \bullet Q_2(r_n)], \quad 1 \leq m \leq M, \end{aligned} \quad (27b)$$

so that (24a,b) simplify to

$$B_{11,mm} \bullet Q_1(r_m) + B_{12,mm} \bullet Q_2(r_m) = Q_{1exc}(r_m), \quad (28a)$$

$$B_{21,mm} \bullet Q_1(r_m) + B_{22,mm} \bullet Q_2(r_m) = Q_{2exc}(r_m). \quad (28b)$$

Hence,

$$C_{11,m} \bullet Q_{1exc}(r_m) + C_{12,m} \bullet Q_{2exc}(r_m) = Q_1(r_m), \quad (28c)$$

$$C_{21,m} \bullet Q_{1exc}(r_m) + C_{22,m} \bullet Q_{2exc}(r_m) = Q_2(r_m), \quad (28d)$$

with

$$C_{11,m} = -[B_{12,mm} \bullet B_{22,mm}^{-1} \bullet B_{21,mm} - B_{11,mm}]^{-1}, \quad (29a)$$

$$\underline{\underline{C}}_{12,m} = [\underline{\underline{B}}_{21,mm} - \underline{\underline{B}}_{22,mm} \bullet \underline{\underline{B}}_{12,mm}^{-1} \bullet \underline{\underline{B}}_{11,mm}]^{-1}, \quad (29b)$$

$$\underline{\underline{C}}_{21,m} = [\underline{\underline{B}}_{12,mm} - \underline{\underline{B}}_{11,mm} \bullet \underline{\underline{B}}_{21,mm}^{-1} \bullet \underline{\underline{B}}_{22,mm}]^{-1}, \quad (29c)$$

$$\underline{\underline{C}}_{22,m} = -[\underline{\underline{B}}_{21,mm} \bullet \underline{\underline{B}}_{11,mm}^{-1} \bullet \underline{\underline{B}}_{12,mm} - \underline{\underline{B}}_{22,mm}]^{-1}. \quad (29d)$$

In keeping with Varadan et al. [15], *rotpole* moments may be defined as

$$\underline{\underline{t}}_{1m} = (i/\omega) \underline{\underline{v}}_m \underline{\underline{W}}_{1eq}(\underline{\underline{r}}_m), \quad (30a)$$

$$\underline{\underline{t}}_{2m} = (i/\omega) \underline{\underline{v}}_m \underline{\underline{W}}_{2eq}(\underline{\underline{r}}_m), \quad (30b)$$

these rotpole moments serving as the Beltrami analogs of the electric dipole moment  $\underline{\underline{p}}_m$  and the magnetic dipole moment  $\underline{\underline{m}}_m$ . It follows from (18c,d) that

$$\underline{\underline{t}}_{1m} = (i/\omega) \underline{\underline{v}}_m [\gamma_1/2k_a] [a_{11} \underline{\underline{Q}}_1(\underline{\underline{r}}_m) + a_{12} \underline{\underline{Q}}_2(\underline{\underline{r}}_m)], \quad (31a)$$

$$\underline{\underline{t}}_{2m} = (i/\omega) \underline{\underline{v}}_m [\gamma_2/2k_a] [a_{21} \underline{\underline{Q}}_1(\underline{\underline{r}}_m) + a_{22} \underline{\underline{Q}}_2(\underline{\underline{r}}_m)], \quad (31b)$$

whence, using (28c,d),

$$\underline{\underline{t}}_{1m} = \underline{\underline{\pi}}_{11,m} \bullet \underline{\underline{Q}}_{1exc}(\underline{\underline{r}}_m) + \underline{\underline{\pi}}_{12,m} \bullet \underline{\underline{Q}}_{2exc}(\underline{\underline{r}}_m), \quad (32a)$$

$$\underline{\underline{t}}_{2m} = \underline{\underline{\pi}}_{21,m} \bullet \underline{\underline{Q}}_{1exc}(\underline{\underline{r}}_m) + \underline{\underline{\pi}}_{22,m} \bullet \underline{\underline{Q}}_{2exc}(\underline{\underline{r}}_m), \quad (32b)$$

with the *rotability* dyadics given as

$$\underline{\underline{\pi}}_{11,m} = (i/\omega) \underline{\underline{v}}_m [\gamma_1/2k_a] [a_{11} \underline{\underline{C}}_{11,m} + a_{12} \underline{\underline{C}}_{21,m}], \quad (32c)$$

$$\underline{\underline{\pi}}_{12,m} = (i/\omega) \underline{\underline{v}}_m [\gamma_1/2k_a] [a_{11} \underline{\underline{C}}_{12,m} + a_{12} \underline{\underline{C}}_{22,m}], \quad (32d)$$

$$\underline{\underline{\pi}}_{21,m} = (i/\omega) \underline{\underline{v}}_m [\gamma_2/2k_a] [a_{21} \underline{\underline{C}}_{11,m} + a_{22} \underline{\underline{C}}_{21,m}], \quad (32e)$$

$$\underline{\underline{\pi}}_{22,m} = (i/\omega) \underline{\underline{v}}_m [\gamma_2/2k_a] [a_{21} \underline{\underline{C}}_{12,m} + a_{22} \underline{\underline{C}}_{22,m}]. \quad (32f)$$

### 4.3 The coupled dipole method (CDM)

The expressions for the rotpole moments derived in the preceding section can be used for another numerical approach; namely the coupled dipole method. This method has been variously known as the discrete dipole method and the Purcell-Pennypacker approach and, as the third name suggests, was heuristically derived by Purcell and Pennypacker in 1973 [16]. It has gained ground for the study of scattering in recent years [17-19]. The basis of this method lies in assuming that each subregion can be thought of in dipolar terms. The total

scattering can then be calculated by summing the radiation due to the dipoles or, as in this case, rotipoles.

In a slightly different form, expressions for the exciting Beltrami fields (27a,b), can be written as

$$\mathbf{Q}_{1exc}(\mathbf{r}_m) = \mathbf{Q}_{1inc}(\mathbf{r}_m) + (\gamma_1 + \gamma_2) \sum_{n=1,2,\dots,M; n \neq m} \{v_n \underline{\underline{G}}_{1mn} \bullet \mathbf{W}_{1eq}(\mathbf{r}_n)\}, \quad (33a)$$

$$\mathbf{Q}_{2exc}(\mathbf{r}_m) = \mathbf{Q}_{2inc}(\mathbf{r}_m) - (\gamma_1 + \gamma_2) \sum_{n=1,2,\dots,M; n \neq m} \{v_n \underline{\underline{G}}_{2mn} \bullet \mathbf{W}_{2eq}(\mathbf{r}_n)\}. \quad (33b)$$

Substitution for the equivalent Beltrami source densities by the rotipoles leads to

$$\mathbf{Q}_{1exc}(\mathbf{r}_m) = \mathbf{Q}_{1inc}(\mathbf{r}_m) - i\omega(\gamma_1 + \gamma_2) \sum_{n=1,2,\dots,M; n \neq m} \{\underline{\underline{G}}_{1mn} \bullet \mathbf{t}_{1n}\}, \quad (34a)$$

$$\mathbf{Q}_{2exc}(\mathbf{r}_m) = \mathbf{Q}_{2inc}(\mathbf{r}_m) + i\omega(\gamma_1 + \gamma_2) \sum_{n=1,2,\dots,M; n \neq m} \{\underline{\underline{G}}_{2mn} \bullet \mathbf{t}_{2n}\}. \quad (34b)$$

Finally, the use of (32a,b) yields the CDM equations

$$\begin{aligned} \mathbf{Q}_{1inc}(\mathbf{r}_m) = \sum_{n=1,2,\dots,M} [\underline{\underline{D}}_{11,mn} \bullet \mathbf{Q}_{1exc}(\mathbf{r}_n) \\ + \underline{\underline{D}}_{12,mn} \bullet \mathbf{Q}_{2exc}(\mathbf{r}_n)], \quad 1 \leq m \leq M, \end{aligned} \quad (35a)$$

$$\begin{aligned} \mathbf{Q}_{2inc}(\mathbf{r}_m) = \sum_{n=1,2,\dots,M} [\underline{\underline{D}}_{21,mn} \bullet \mathbf{Q}_{1exc}(\mathbf{r}_n) \\ + \underline{\underline{D}}_{22,mn} \bullet \mathbf{Q}_{2exc}(\mathbf{r}_n)], \quad 1 \leq m \leq M, \end{aligned} \quad (35b)$$

where the  $6M$  unknowns are the components of  $\mathbf{Q}_{1exc}(\mathbf{r}_m)$  and  $\mathbf{Q}_{2exc}(\mathbf{r}_m)$ , and the dyadics involved are given as

$$\underline{\underline{D}}_{11,mn} = \underline{\underline{I}}\delta_{mn} + i\omega(1 - \delta_{mn})(\gamma_1 + \gamma_2) \underline{\underline{G}}_{1mn} \bullet \underline{\underline{\pi}}_{11,n}, \quad (36a)$$

$$\underline{\underline{D}}_{12,mn} = i\omega(1 - \delta_{mn})(\gamma_1 + \gamma_2) \underline{\underline{G}}_{1mn} \bullet \underline{\underline{\pi}}_{12,n}, \quad (36b)$$

$$\underline{\underline{D}}_{21,mn} = -i\omega(1 - \delta_{mn})(\gamma_1 + \gamma_2) \underline{\underline{G}}_{2mn} \bullet \underline{\underline{\pi}}_{21,n}, \quad (36c)$$

$$\underline{\underline{D}}_{22,mn} = \underline{\underline{I}}\delta_{mn} - i\omega(1 - \delta_{mn})(\gamma_1 + \gamma_2) \underline{\underline{G}}_{2mn} \bullet \underline{\underline{\pi}}_{22,n}. \quad (36d)$$

The scattered fields in the CDM can be computed after the calculated excitation fields have been used to determine the dipole and the rotipole moments. Thus,

$$\mathbf{Q}_{1sca}(\mathbf{r}) = -i\omega(\gamma_1 + \gamma_2) \sum_{n=1,2,\dots,M} \{\underline{\underline{G}}_1(\mathbf{r}, \mathbf{r}_n) \bullet \mathbf{t}_{1n}\}, \quad \mathbf{r} \in V_{ext}, \quad (34a)$$

$$\mathbf{Q}_{2sca}(\mathbf{r}) = i\omega(\gamma_1 + \gamma_2) \sum_{n=1,2,\dots,M} \{\underline{\underline{G}}_2(\mathbf{r}, \mathbf{r}_n) \bullet \mathbf{t}_{2n}\}, \quad \mathbf{r} \in V_{ext}. \quad (34b)$$

## 5. Concluding Remarks

To summarize, we have considered scattering of Beltrami fields from a chiral inclusion embedded in a chiral host. In our treatment, we replaced the scatterer by equivalent current densities and then volume integral equations were obtained for the scattered fields. These equations were then reduced to algebraic equations and two algorithms – method of moments and the coupled dipole method – were developed. These equations can be easily solved using standard matrix manipulations provided they are not algorithmically singular. The expressions derived here are much more simpler in form than the corresponding ones for the electric and magnetic fields. It is also important to note that the scattering problem is solved for an arbitrary shape of the inclusion. Finally, the problem solved in this paper forms a part of Ref. 20 and 21.

## References

1. Bohren, C. F., 1974, "Light scattering by an optically active sphere," *Chem. Phys. Lett.* **29**, 458-462.
2. Silberstein, L., 1907, "Elektromagnetische grundgleichungen in bivectorieller behandlung," *Ann. Phys. Leipzig* **22**, 579-587.
3. McKelvey, J. P., 1990, "The case of the curious curl," *Amer. J. Phys.* **58**, 306-310.
4. Yoshida, Z., 1991, "Helicity of waves propagating in a plasma," *J. Plasma Phys.* **45**, 481-488.
5. Varadan, V. K., Lakhtakia, A., and Varadan, V. V., 1987, "A comment on the solution of the equation  $\nabla \times \mathbf{a} = \mathbf{ka}$ ," *J. Phys. A: Math. Gen.* **20**, 2649-2650.
6. Weiglhofer, W. S., 1989, "A simple and straightforward derivation of the dyadic Green's function of an isotropic chiral medium," *Arch. Elektron. Über.* **43**, 51-52.

7. Lakhtakia, A., 1991, "Alternative derivation of the infinite-medium dyadic Green's function for isotropic chiral media," *Arch. Elektron. Über.* **45**, 323-324.
8. Shanker, B., 1993, "On dyadic Green's function for Beltrami fields," *Int. J. Infrared Milim. Waves* **14**, 1077-1081.
9. Lakhtakia, A., 1992, "Strong and weak forms of the method of moments and the coupled dipole method for scattering of time-harmonic electromagnetic fields," *Int. J. Mod. Phys. C* **3**, 583-603.
10. Lakhtakia, A., 1991, "First order characterization of electromagnetic fields in isotropic chiral media," *Arch. Elektron. Über.* **45**, 57-59.
11. Jackson, J. D., 1975, *Classical Electrodynamics*, Wiley, New York.
12. Harrington, R. F., 1968, *Field Computations by Moment Methods*, Macmillan, New York.
13. Miller, E. K., 1988, "A selective survey of computational electromagnetics," *IEEE Trans. Antennas Propagat.* **36**, 1281-1305.
14. Wang, J. J. H., 1991, *Generalized Moment Methods in Electromagnetics*, Wiley, New York.
15. Varadan, V. V., Lakhtakia, A., and Varadan, V. K., 1991, "Microscopic circular polarizabilities (rotabilities) and the macroscopic properties of chiral media," *Radio Sci.* **26**, 511-516.
16. Purcell, E. M., and Pennypacker, C. R., 1973, "Scattering and absorption by nonspherical dielectric grains," *Astrophys. J.* **186**, 705-714.
17. Singham, S. B., and Bohren, C. F., 1988, "Light scattering by an arbitrary particle: the scattering-order formulation of the coupled-dipole method," *J. Opt. Soc. Am. A* **5**, 1867-1872.

18. Lakhtakia, A., 1992, "General theory of the Purcell-Pennypacker scattering approach and its extension to bianisotropic scatterers," *Astrophys. J.* **394**, 494-499.
19. Dungey, C. E., and Bohren, C. F., 1991, "Light scattering by nonspherical particles: a refinement to the coupled-dipole method," *J. Opt. Soc. Am. A* **8**, 81-87.
20. Shanker, B., and Lakhtakia, A., 1993, "Extended Maxwell Garnett model for chiral-in-chiral composites," *J. Phys. D. Appl. Phys.* **0**, 00-00; in press.
21. Lakhtakia, A., and Shanker, B., 1993, "Beltrami fields within continuous source regions, volume integral equations, scattering algorithms and the extended Maxwell-Garnett model," *Int. J. Appl. Electromag. Matls.* **4**, 65-82.

# Forced Lattice Vibrations

Percy Deift and Thomas Kriecherbauer

Courant Institute  
New York University

Stephanos Venakides  
Duke University

## 1 Introduction.

We study shock waves in dispersive semi-infinite particle chains generated by an imposed motion on the first particle. We assume the imposed velocity of the first particle to be a periodic function of time with positive mean value. The first particle moves toward the others compressing the chain. For a chain with nearest neighbor interactions the motion is described by:

$$(1.1a) \quad \frac{d^2}{dt^2}x_n = F(x_{n+1} - x_n) - F(x_n - x_{n-1}).$$
$$n = 1, 2, 3, \dots,$$

The initial positions and velocities are given by:

$$(1.1b) \quad x_n(0) = 0, \quad \dot{x}_n(0) = 0 \quad n = 1, 2, \dots,$$

and the forcing velocity is given by

$$(1.1c) \quad \dot{x}_0(t) = 2a + f(t) \quad t \geq 0, \quad a > 0,$$

where  $x_n(t)$  is the deviation of the  $n^{\text{th}}$  particle from an initial rest position and  $f(t)$  is a zero-mean periodic function of time.

We first outline the case of constant forcing ( $f(t) \equiv 0$ ). Holian and Straub [4] took various force laws  $F(\cdot)$  associated with the names Lennard-Jones, Morse, and Toda, and investigated the limit  $t \rightarrow \infty$ . Through numerical experimentation, they not only confirmed an earlier finding of von-Neumann [8] that the shock wave is reflected from the origin with large oscillations behind the shock, but discovered a striking new phenomenon, the existence of a critical shock strength  $a_{\text{crit}}$  with the following property:

- (i) For  $a < a_{\text{crit}}$  the solution of the shock problem is oscillatory in the region  $s_2 t < n < s_1 t$  while it tends to zero in the region  $0 < n < s_2 t$ , as  $t$  tends to infinity. Here  $s_1$  is the shock speed and  $s_2$  a second kind of speed; both  $s_1$  and  $s_2$  are functions of  $a$ .
- (ii) For  $a > a_{\text{crit}}$  the solution is oscillatory everywhere behind the shock, i.e. for  $0 < n < s_1 t$ . Yet, the quality of the oscillations is different in the inner region: it is periodic in time and binary in space i.e. neighboring particles are moving with opposite velocities relative to the forcing particle except in the region of the zeroth particle itself where motion is somewhat more complicated. This is precisely explained in [7].

The existence of a critical forcing velocity is not yet generally understood. In the special case of the Toda chain (also referred to as the Toda lattice) which has exponential force law  $F(r) = -e^{-r}$ , Holian, Flaschka and McLaughlin [3] utilized complete integrability to analyze the von-Neumann problem and derived the critical speed  $a = 1$  and the shock speed  $s_1(a)$ . Indeed, by translating the coordinates so that the particle  $x_0$  remains fixed and reflecting the chain about this particle, one imbeds the von-Neumann problem, which is nonautonomous, into an autonomous initial value problem for a doubly infinite chain ( $-\infty < n < \infty$ ). The particles in this chain still satisfy equations (1.1a). They have zero initial deviations and have initial velocities given by  $-2a \operatorname{sgn}$



where  $n$  is the particle index. In the case of the Toda chain the new problem is completely integrable.

In the integrable problem, criticality arises from the fact that the continuous spectrum of the Lax operator, given by the set  $[-a-1, -a+1] \cup [a-1, a+1]$  consists of a single interval in the subcritical case  $0 < a < 1$  and of two disjoint intervals in the supercritical case  $a > 1$ .

Venakides, Deift and Oba [7] studied the supercritical von-Neumann problem for the Toda chain by analyzing the long time behavior of the  $\tau$ -function of the imbedding integrable problem on  $-\infty < n < \infty$ . They showed the emergence of oscillations whose structure they obtained in detail and derived the speeds  $s_1$  and  $s_2$ . (They denote these speeds by  $N_{max}$  and  $N_{min}$  respectively.) Using the same technique, Kamvissis [5] derived the oscillatory structure in the critical and subcritical case.

The striking result in [7] and [5] is that the residual state of the chain i.e. the state of motion in the region  $|n| < s_2 t$ , which each particle eventually enters, is described by an algebraic/geometric type solution whose Lax operator has exactly the same spectrum as the Lax operator of the original integrable shock problem. In other words, (a) the problem remains isospectral even after a spatially non-uniform limiting process ( $t \rightarrow \infty$ ) eliminates the region  $s_2 t < |n|$  by pushing it to infinity and (b) in the residual state, there are no more degrees of freedom except those dictated by the spectrum: the number of degrees of freedom equals the number of spectral gaps which in this case is effectively equal to one.

When the forcing velocity  $f(t)$  is no longer constant, numerical experiments (see below) indicate that the residual state is again described, in the region  $n < const.t$ , by an algebraic/geometric type solution, but now the number of gaps, and hence the number of degrees of freedom, may be greater than one.

Unfortunately, the process of imbedding the semi-infinite chain into an integrable chain does not work when the forcing velocity is time dependent. Although the problem may still be integrable, as one might be led to suspect

by the recent work of Fokas and Its [FI] on initial-boundary value integrable problems, no linearizing transformation has yet been found. Faced with the inability to reflect the problem we ask ourselves: How does the spectrum of the residual state manifest itself in the original forced semi-infinite Toda chain?

The evolution of the Lax operator of the semi-infinite chain is not isospectral; the Lax-pair equation has a rank-one perturbing term which contains the forcing. The initial ( $t = 0$ ) continuous spectrum of the Lax-operator for the semi-infinite chain consists of the single interval  $[a-1, a+1]$ . The spectral evolution, studied numerically, is as follows. When the forcing velocity is constant, eigenvalues are emitted from the lowest point of the continuous spectrum at a constant rate and move to the left filling the band  $[-a-1, -a+1]$  (actually the band  $[-a-1, a-1]$  when  $a < 1$ ) in the limit  $t \rightarrow \infty$ , which we also consider as continuous spectrum. The limiting "continuous" spectrum of the Lax operator is identical to the continuous spectrum of the residual state. The emission of eigenvalues from the continuous spectrum was first observed by Kaup and Neuberger [6], who make an approximate calculation of the eigenvalue birth rate. When the forcing velocity contains a periodic component ( $f \neq 0$ ), the emitted eigenvalues may fill more than one band corresponding to a residual state that is a multiphase wave.

In our analysis of the periodically forced problem we achieve the following:

- (a) We derive a closed system for the evolution of the scattering data, including the evolution of eigenvalues.
- (b) We make the Ansatz that the eigenvalues asymptotically cluster onto a finite set of bands. (This is confirmed by numerical experiment in the subcritical case.) We then take the continuum limit of the eigenvalue evolution equations to derive an integral equation for the asymptotic ( $t \rightarrow \infty$ ) spectral density of these bands. We cannot yet determine the number, and the endpoints of the bands (if  $2g$  numbers are needed we only have  $g$  relations). However if we assume the endpoints given, then

we can correctly calculate the corresponding spectral densities.

- (c) In the subcritical case of the Toda chain, and when the periodic component of the forcing is small, we rigorously construct, under minimal assumptions, solutions to the chain equations that are valid for  $n > 0$  and  $-\infty < t < \infty$  and in which the velocity of the zeroth particle is a given time-periodic function. The solutions are multiphase waves away from the boundary and they connect to the forcing function through a boundary layer. When the frequency exceeds a threshold value, there is no wave penetration into the chain, there is only a boundary layer that decays exponentially with the particle index. As the frequency decreases, more phases are generated according to a precise formula. If the number of the wave-phases is  $g$ , the location of the midpoints of the  $g$  spectral gaps is determined by the frequency of the driver. Their widths as well as the corresponding phase shifts ( $2g$  pieces of information) are determined by the first  $g$  Fourier coefficients of the driver ( $2g$  pieces of information, since these coefficients are in general complex). The remaining part of the information, i.e. the remaining Fourier coefficients of the driver determine the boundary layer. The calculation involves a Liapunov Schmidt decomposition, the definition of an appropriate family of norms and the use of the implicit function theorem. We can carry through the above calculation for a general nonintegrable chain, if we restrict ourselves to single phase waves. This is equivalent to requiring that the frequency of the driver be sufficiently large. We have not yet been able to construct general nonintegrable chain multiphase waves due to problems with small divisors.

In this presentation, we focus on the case in which the periodic component of the driver is not small, i.e. as described in points (a) and (b) above.

## 2 The Evolution Equations

We use the Flaschka variables

$$(2.1) \quad a_n = -\frac{\dot{s}_n}{2}, \quad b_n = \frac{1}{2}e^{(s_n - s_{n+1})/2}, \quad n = 0, 1, 2, \dots$$

and we note that the function  $a_0(t)$  is the given time-periodic forcing function. Equations (1.1a) for the semi-infinite chain with  $F(r) = e^{-r}$  (Toda) are easily reduced to the perturbed Lax pair equation:

$$(2.2) \quad \frac{dM}{dt} + MB - BM = -\rho(t)P$$

where  $M$  is the tridiagonal matrix

$$M = \begin{bmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & 0 & \\ & b_2 & a_3 & b_3 & \\ 0 & & & \ddots & \ddots \end{bmatrix},$$

$B$  is the antisymmetric tridiagonal matrix given by

$$B = \begin{bmatrix} 0 & b_1 & & & \\ -b_1 & 0 & b_2 & & \\ 0 & -b_2 & 0 & b_3 & \\ & & & \ddots & \ddots \end{bmatrix},$$

$P$  is the rank-one matrix given by:

$$P_{ij} = \begin{cases} 1 & \text{if } i = j = 1 \\ 0 & \text{otherwise} \end{cases}$$

and  $\rho(t)$  is the function:

$$\rho(t) = 2b_0^2(t) = 2b_1^2(t) - \dot{a}_1(t), \quad \cdot = \frac{d}{dt}.$$

Strictly speaking, the matrices  $M$  and  $B$  are semi-infinite. However, truncating the chain at some particle of very large index  $N$  makes the matrices  $M$  and  $B$  finite. The disturbance in the chain caused by the truncation, travels

essentially with finite velocity. Only exponentially small effects display infinite speed. Thus, a large part of the chain, say the left half, does not essentially feel the truncation until a time  $t = O(N)$ . Our analysis is in the regime  $1 \ll t \ll N$  and  $n \ll N$  of a finite chain.

Our strategy is to derive evolution equations for:

- (a) the eigenvalues  $\lambda_j$  of the truncated matrix  $M$ .
- (b) The first entry  $f_j$  of the  $j^{\text{th}}$  eigenvector of  $M$  ( $j = 1, \dots, N$ ) when it is normalized to have Euclidean length equal to one.

Given the  $N$  eigenvalues  $\lambda_i$  and the first eigenvector entries  $f_i$ , we can reconstruct the matrix  $M$  [7].

**Theorem 2.3.** The evolution of the  $\lambda_j$ 's and  $f_j$ 's is given by:

$$\begin{cases} \frac{1}{2} \frac{d}{dt} \ln(-\dot{\lambda}_j) = \lambda_j - a_0(t) + \sum_{i \neq j}^N \frac{\dot{\lambda}_i}{\lambda_j - \lambda_i}, j = 1, \dots, N. \\ f_j^2 = \frac{-\dot{\lambda}_j}{\rho} \end{cases}$$

where  $\rho = 2b_0^2(t) = -\sum_{i=1}^N \dot{\lambda}_i$ , and a dot indicates a derivative with respect to time.

The initial values  $\lambda_i(0)$  are the eigenvalues of  $M$  at  $t = 0$  while the initial values  $\dot{\lambda}_i(0)$  are given by

$$\dot{\lambda}_i(0) = -2b_0^2(0)f_i^2(0).$$

**Proof:** Let  $\Lambda$  be the diagonal matrix of the eigenvalues  $\lambda_j$  of  $M$  and let  $\Psi$  be the matrix whose  $j^{\text{th}}$  column is the normalized eigenvector of  $M$  corresponding to the eigenvalue  $\lambda_j$ . We have:

$$(2.4a) \quad M\Psi = \Psi\Lambda,$$

Let

$$(2.4b) \quad \Phi = \dot{\Psi} - B\Psi \quad \text{where} \quad \dot{\phantom{x}} = \frac{d}{dt},$$

Utilizing equations (2.4a) and (2.4b), we easily calculate:

$$(2.4c) \quad M\dot{\Phi} - \dot{\Phi}\Lambda = \Psi\dot{\Lambda} + \rho P\Psi.$$

We now define the matrix  $A = (a_{ij})$  by the relation:

$$(2.4d) \quad \Phi = \Psi A.$$

We calculate ( $A^T$  is the transpose of  $A$ ):

$$A^T + A = \Psi^T \dot{\Phi} + \dot{\Phi}^T \Psi = \Psi^T (\dot{\Psi} - B\Psi) + (\dot{\Psi}^T - \Psi^T B^T) \Psi = \Psi^T \dot{\Psi} + \dot{\Psi}^T \Psi = \frac{d}{dt}(\Psi^T \Psi) = 0.$$

The last equality is true because the matrix  $\Psi^T \Psi$  is orthogonal. Thus

$$(2.5) \quad A^T + A = 0;$$

i.e.  $A$  is antisymmetric. Using (2.4c) we obtain

$$(2.6) \quad M\dot{\Phi} - \dot{\Phi}\Lambda = M\Psi A - \Psi A\Lambda = \Psi(\Lambda A - A\Lambda).$$

Comparing (2.4c) with (2.6) we obtain easily:

$$(2.7) \quad \dot{\Lambda} = [\Lambda, A] - \rho \Psi^T P \Psi$$

Let  $f^T = (f_1, f_2, \dots, f_N)$  be the first row of  $\Psi$ . Then

$$\Psi^T P \Psi = f f^T$$

where the righthand side is a matrix product. We insert this in (2.7).

$$(2.8) \quad \dot{\Lambda} = [\Lambda, A] - \rho f f^T.$$

Equating the diagonal elements on both sides we obtain:

$$(2.9) \quad \dot{\lambda}_j = -\rho f_j^2, \quad \sum_{j=1}^N \dot{\lambda}_j = -\rho.$$

This proves the second relation in Theorem 2.3.

Off the diagonal in (2.8) we have for  $i \neq j$ :  $\lambda_i a_{ij} - a_{ij} \lambda_j = \rho f_i f_j$ , hence:

$$(2.10) \quad a_{ij} = \frac{\rho f_i f_j}{\lambda_i - \lambda_j} \quad \text{when } i \neq j;$$

$a_{ii} = 0$  by (2.5). We now calculate the evolution of  $f_j$ . By (2.4b):

$$\dot{\Psi} = \Phi + B\Psi = \Psi A + B\Psi.$$

Specializing this to the first row we obtain:  $\dot{f}^T = f^T A + B_{R_1} \Psi$  where  $B_{R_1}$  is the first row of  $B$ .

$$\dot{f}^T = f^T A + (M - a_1 T)_{R_1} \Psi = f^T A + (M \psi)_{R_1} - a_1 f^T = f^T A + (\Psi \Lambda)_{R_1} - a_1 f^T = f^T A + f^T \Lambda - f^T a_1$$

Taking transposes and factoring we obtain:

$$(2.11) \quad \dot{f} = (\Lambda - a_1 I - A)f.$$

We write the antisymmetric matrix  $A$  obtained in (2.10) as

$$(2.12) \quad A = \rho F L F$$

where  $F$  is the diagonal matrix with entries  $f_1, \dots, f_N$  and  $L$  is the matrix  $\left(\frac{1}{\lambda_i - \lambda_j}\right)$ . We obtain  $\dot{f} = (\Lambda - a_1 I - \rho F L F)F$ .

$$(2.13) \quad F^{-1} \dot{f} = F^{-1} \Lambda f - a_1 F^{-1} f - \rho L F f.$$

We then remark that:

$$(i) \quad \rho F f = \rho \begin{pmatrix} f_1^2 \\ f_2^2 \\ \vdots \end{pmatrix} = - \begin{pmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \\ \vdots \end{pmatrix} \text{ by (2.9),}$$

$$(ii) F^{-1}f = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix}, F^{-1}\Lambda = \Lambda F^{-1}.$$

Substituting in (2.13) we obtain

$$(2.14) \quad \frac{f_i}{f_j} = \lambda_j - a_1 + \sum_{i=1}^N \frac{\lambda_i}{\lambda_j - \lambda_i}.$$

The evolution of the  $\lambda_i$ 's in (2.3) is finally obtained by eliminating  $f_j$  between (2.14) and (2.9) and using the expression for  $\rho$  given in (2.2).

### 3 Numerical Results

In Figures 1a and 1b respectively, we show the position and velocity profile of the chain when a constant subcritical velocity  $\dot{x}(t) = \dot{x}_0(t) = 1$  is imposed on the leading particle. In Figures 2a and 2b, we display the same information but now for a supercritical driver  $\dot{x}(t) = \dot{x}_0(t) = 4$ . In this case one observes residual binary oscillations in a neighborhood of the forcing particle. The length (=number of particles) of the chain over which oscillations occur is given asymptotically by  $s_1 t$  while the length of the chain that displays binary oscillations equals  $s_2 t$ . Both speeds  $s_1$  and  $s_2$  are calculated in [7].

In Figures 3 through 8 we present the asymptotic (large  $t$ ) position profile and the spectral profile of the lattice when a zero-mean, time-periodic perturbation is added to a subcritical forcing velocity. We vary the frequency and the amplitude. We have chosen three frequencies and two amplitudes. We label the latter as "large amplitude" and "small amplitude".

We observe that when the frequency is large enough, the oscillations do not penetrate the chain (see Figures 3a and 4a). The continuous spectrum emits eigenvalues that fill up the interval  $(-1.5, -0.5)$  as is seen in Figures 3b and 4b. Asymptotically the spectrum changes from the interval  $(-0.5, 1.5)$  to the interval  $(-1.5, 1.5)$ . The new spectrum corresponds to the residual compressed state of the chain that results from the constant part of the forcing.



The phenomenon of non-penetration of the chain by high-frequencies is easily understood in the case of a linear chain where the threshold frequency can be directly calculated. The phenomenon persists in the nonlinear case. When the amplitude of the driver is small, there is again a threshold frequency which we can calculate. For larger amplitudes and for forcing velocities of the form  $\dot{x}_0(t) = 2a + f(\omega t)$ , where  $f$  is  $2\pi$ -periodic and  $\omega > 0$ , it is not yet possible to compute the threshold frequency  $\omega = \omega_0$  below which penetration occurs. In this large amplitude situation,  $\omega_0$  will depend on the shape of  $f$ .

When the forcing frequency and waveshape are such that oscillations begin to arise in the chain, a dramatic change occurs in the spectrum. The emitted eigenvalues cluster in bands as is seen in Figures 5b, 6b, 7b, and 8b. If the driver is of the form  $\dot{x}_0(t) = 2a + \epsilon f(t)$ ,  $0 < a < 1$ ,  $\epsilon$  small, we can verify numerically that the residual state of the chain at  $t = \infty$  is described precisely by solutions of the type constructed in section 1(c), which, therefore, constitute the set of attractors for the forced chain (1.1). We believe, consistent with the numerical evidence, that the same situation occurs when  $\epsilon$  is large. However, we have not yet been able to construct solutions analogous to the ones of section 1(c) in this case. The technical difficulty lies in connecting the boundary forcing at  $n = 0$ , through a boundary layer, to the  $g$ -gap algebraic/geometric solution at large  $n$ . We note again that as the frequency of the driver decreases, more wave-phases are activated and correspondingly more gaps appear in the spectrum.

In Figures 5c, and 7c, we plot the (integrated) spectral density, i.e. the number of eigenvalues below  $\lambda$  (=spectral parameter), divided by time. In Figures 5d, and 7d, we plot the same densities as predicted by our theory that is based on taking the continuum limit of the eigenvalue evolution equations of Theorem 2.3. Our theory is not yet complete. We cannot predict the endpoints of the band/gap spectral structure. To derive the predicted densities we had to use the numerically obtained values for the endpoints.

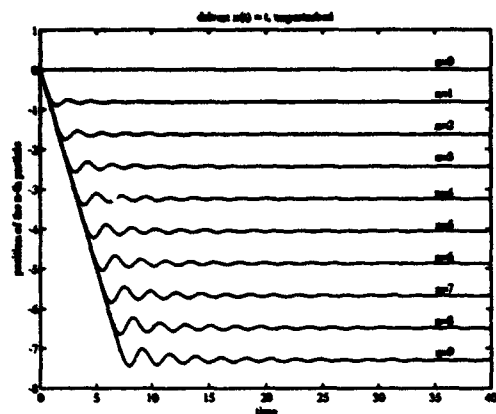


Figure 1(a)

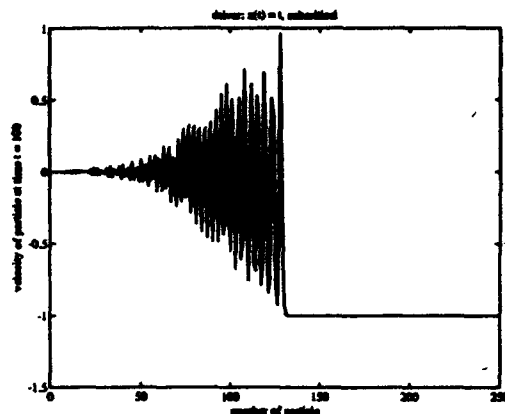


Figure 1(b)

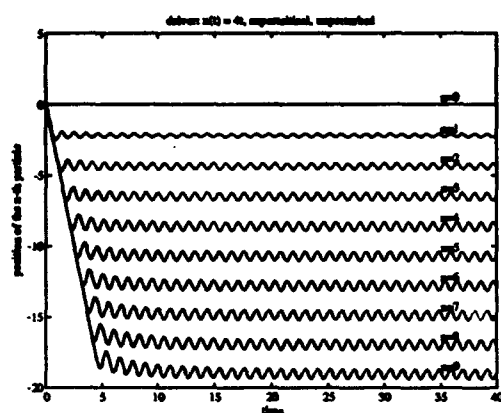


Figure 2(a)

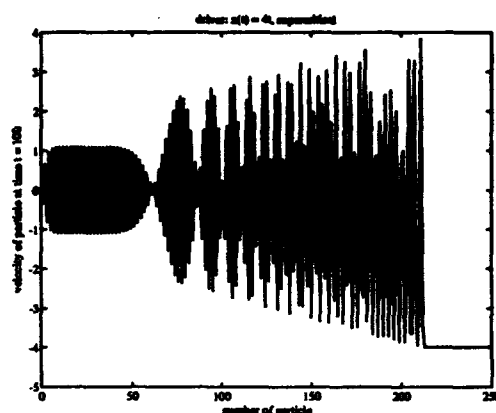


Figure 2(b)

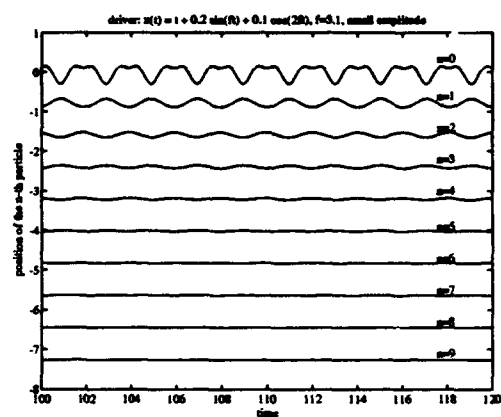


Figure 3(a)

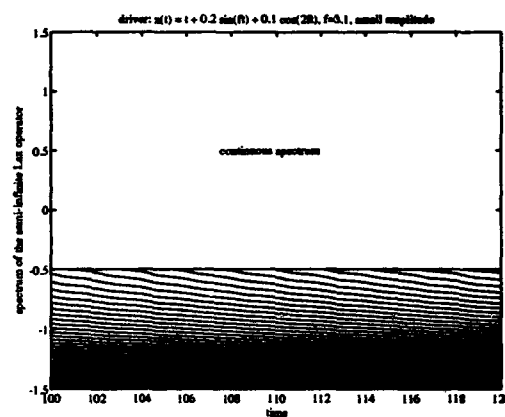


Figure 3(b)

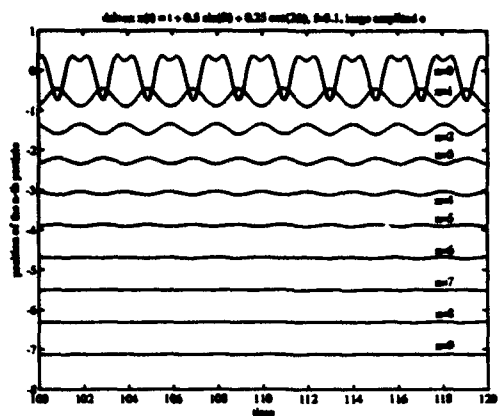


Figure 4(a)

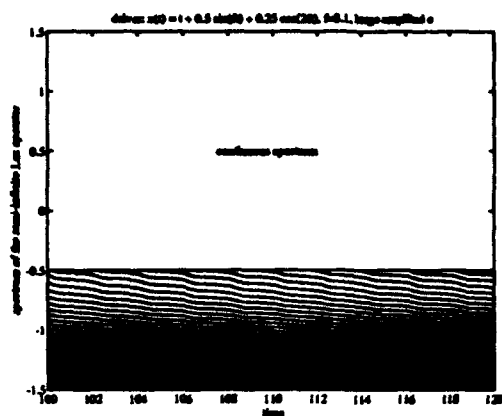


Figure 4(b)

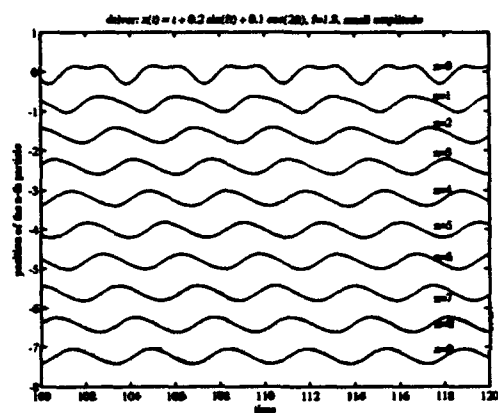


Figure 5(a)

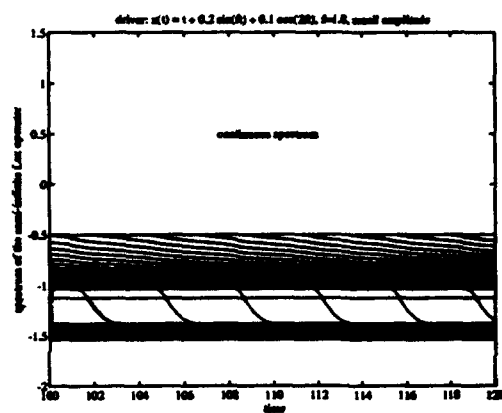


Figure 5(b)

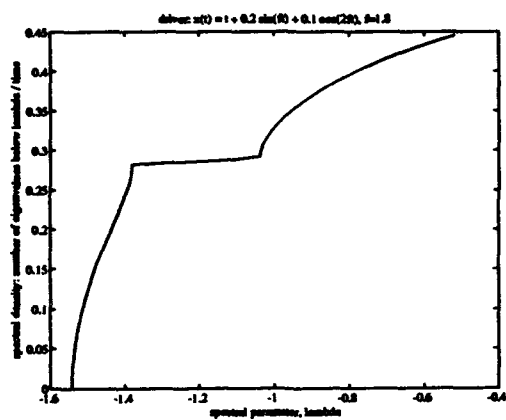


Figure 5(c)

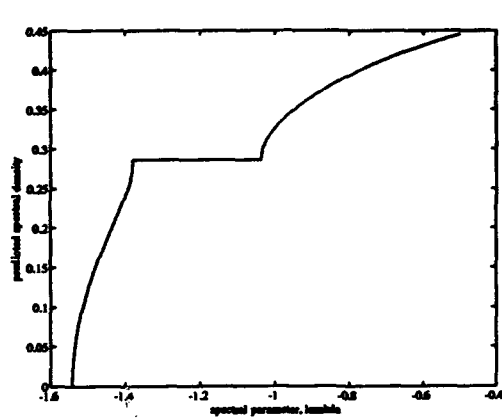


Figure 5(d)

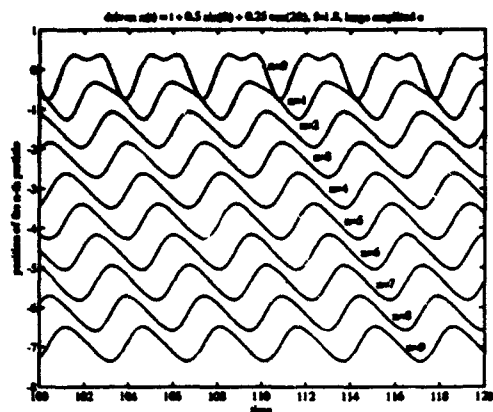


Figure 6(a)

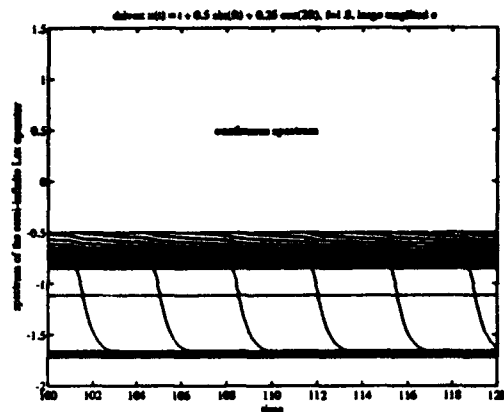


Figure 6(b)

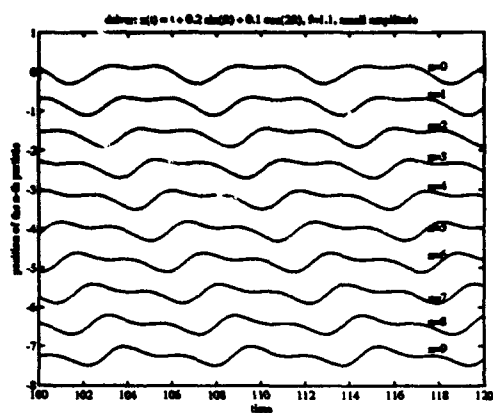


Figure 7(a)

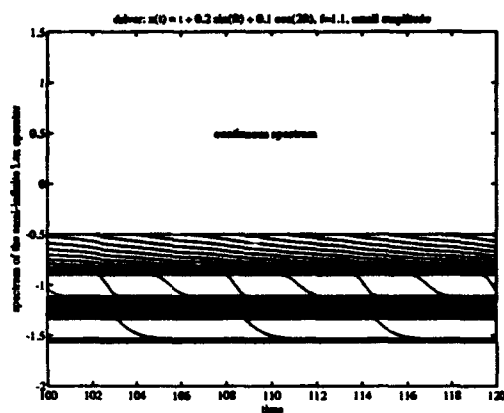


Figure 7(b)

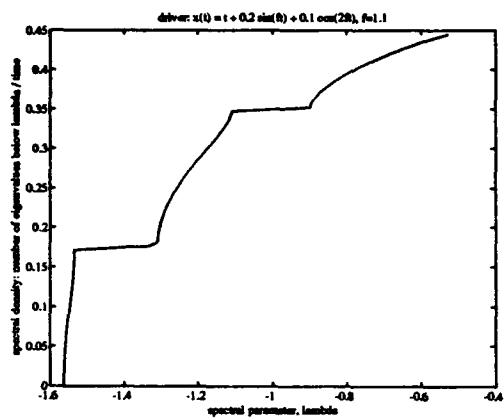


Figure 7(c)

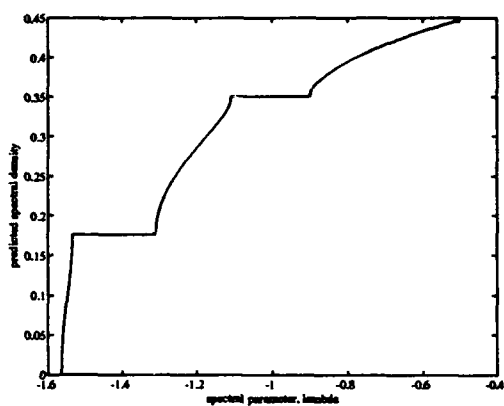


Figure 7(d)

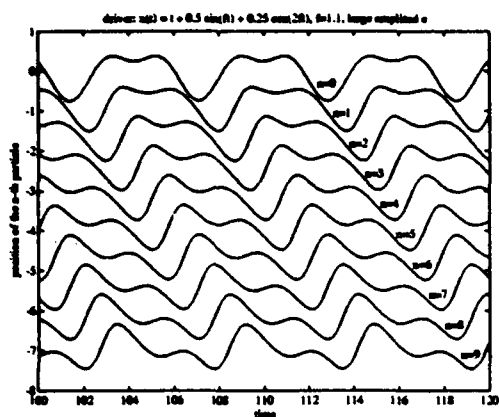


Figure 8(a)

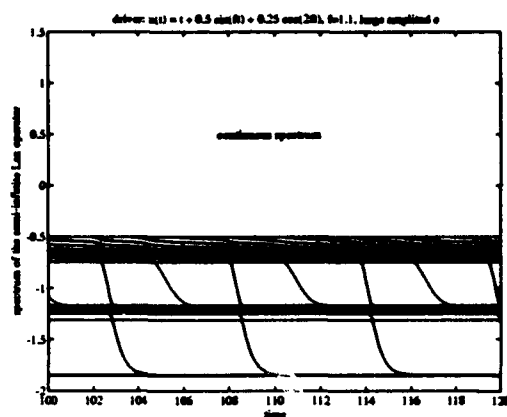


Figure 8(b)

## 4 The Continuum Limit

We outline our main result in taking the continuum limit of the eigenvalue evolution equations of Theorem 2.3. It is clear numerically that eigenvalues are emitted from the continuous spectrum at a rate that is constant if averaged over a large period of time. It therefore is assumed that the number of eigenvalues in an interval grows linearly with time. The density function  $\psi(\lambda)$  is defined as the difference between the asymptotic density of eigenvalues at time  $t$  and at time zero normalized by dividing by the time  $t$ . In other words, the asymptotic (large  $t$ ) number of eigenvalues in an interval  $(\lambda_1, \lambda_2)$  minus the number of eigenvalues in the same interval at time  $t = 0$  is given to leading order by the integral of  $\psi$  over the interval, multiplied by  $t$ . We derive the following continuum limit of the eigenvalue dynamics:

$$(4.1) \quad \lambda - \langle a_0 \rangle - \frac{1}{\pi} \int_{-\infty}^{\infty} \log|\lambda - \mu| \psi(\mu) d\mu = 0 \quad \text{when } \lambda \text{ is in } B,$$

$$\psi(\lambda) = 0 \quad \text{when } \lambda \text{ is not in } B.$$

$$(4.2) \quad \int_{-\infty}^{\infty} \psi(\lambda) d\lambda = 0 \quad (\text{conservation of eigenvalues}).$$

In these formulae  $\langle a_0 \rangle$  is the mean value of the given velocity  $a_0(t)$  and the set  $B$  is the support of the density function  $\psi$ . The set  $B$  is one of the principal unknowns of the problem. The density function  $\psi$  is required to be nonnegative in regions in which the original ( $t = 0$ ) density of eigenvalues is zero.

We utilize the fact that the derivative of the integral kernel in (4.1) with respect to  $\lambda$  is the integral kernel of the Hilbert transform, to reduce these equations to a Riemann-Hilbert problem, which we can solve if we have some additional information on the set  $B$ . This additional information could be gained if we could keep a higher order term as we pass to the continuum limit in our derivation of equations (4.1) and (4.2), which we have not yet succeeded in doing. However, once  $B$  is known from the numerical experiments, we can solve the system (4.1)-(4.2) for  $\psi$ . Integrating  $\psi$ , we obtain the spectral densities of Figures 5d and 7d, which compare favorably with the spectral densities computed numerically directly from the data and displayed in Figures 5c and 7c.

## 5 References

1. P. Deift, M. McDonald, S. Venakides, Renormalization of the  $\tau$ -function of the Toda Lattice, in preparation.
2. A.S. Fokas, A.R. Its, The Linearization of the Initial Boundary Value Problem of the Nonlinear Schroedinger Equation, preprint 1993.
3. B.L. Holian, H. Flaschka, D.W. McLaughlin, Shock Waves in the Toda Lattice: Analysis, Phys. Rev. A. 24 (5)(1981),2595-2623.
4. B.L. Holian, G.K. Straub, Molecular Dynamics of Shock Waves in One-dimensional Chains, Phys. Rev. B. 18 (1978), 1593.
5. S. Kamvissis, On the Long -Time Behavior of the Doubly Infinite Toda Chain Under Shock Initial Data, Dissertation, NYU, 1991.

6. D.J. Kaup, D.H. Neuberger, The Soliton Birth Rate in the Forced Toda Lattice., J. Math. Phys. 25 (1984), 282-284.
7. S. Venakides, P. Deift, R. Oba, The Toda Shock Problem, CPAM. 44 (1991), 1171-1242.
8. J. von Neumann, R.D. Richtmyer, A method for the Numerical Calculation of Hydrodynamic Shocks, J. Appl. Phys. 21 (1950), 7-19.

# Asymptotic Model for Deflagration-to-Detonation Transition in Reactive Two-Phase Flow

Pedro F. Embid\*

## Abstract

The Deflagration-to-Detonation Transition (DDT) in energetic granular materials is a complex multiphase phenomena involving many thermal and mechanical effects among the solid and gas phases. Baer and Nunziato introduced a continuum system of equations in their study of DDT in reactive two-phase flows. This system becomes resonant when the compaction wave speed of the solid equals one of the gas-acoustic wave speeds. Combining high-energy activation energy asymptotics with nonlinear geometrical optics, we develop an asymptotic model for the resonant wave interaction of a fast moving burning front and one of the gas-acoustic waves. This model is capable of predicting in qualitative fashion several of the scenarios documented for DDT through large scale computations. In addition, the various scenarios occur for different choices of parameters in the asymptotic model, which have direct physical interpretation in terms of the asymptotic procedure.

## 1 Introduction

Understanding the process of Deflagration-to-Detonation Transition (DDT) is a key safety issue for the industry and the military [4]. The problem is particularly complex in reactive multi-phase media. However, there are several identifiable stages in DDT [1, 4]. In the initial conductive stage the flame is propagated mainly by heat diffusion. Next there is the convective stage where the flame rapidly grows and accelerates through heat and momentum transfer between the solid reactant and the hot product gases. This is followed by a compressive stage characterized by further acceleration and the development of a shock in the solid phase. Finally, the wave continues to grow to a fully developed detonation wave. This description indicates that heat and momentum transfer between the solid and gas phases are important mechanisms in DDT. Among these transfer effects one has the preheating of the solid explosive by the generated hot gases, the formation of zones of high compaction in the solid and choking of the gas flow in the convective stage, and load transfer leading to shock formation in the solid in the compressive stage.

Recently, Baer, Nunziato, and their collaborators [1, 2, 3] developed a complex system of equations for describing transition to detonation in reactive granular materials. Their

---

\*Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, partially supported by ARO DAALO3-91-G-0186, NSF DMS-9103551 and Sandia Nat. Labs. contract no. AG-8346



approach involves the rational continuum theory of mixtures and a systematic use of the second law of thermodynamics. The resulting system of nonlinear equations involves fluid mechanics for the gaseous phase and solid mechanics for the grains coupled together through an equation for the volume fraction of the solid as well as source terms involving the bulk effects of surface chemistry and compaction.

Numerical integration of this system of equations yields excellent qualitative and quantitative agreement with the available experimental data for various types of secondary explosives, such as CP and HMX. This data involves the trajectory of the flame front, run distance to detonation and the time to detonation for various initial values of the precompacted granular bed [1, 2, 3].

However, an important theoretical problem consists in identifying key physical mechanisms that are responsible for the formation of hot spots and the transition to detonation in multi-phase flows. Rather than attempt a direct study of the formidable system for reactive multi-phase flow, we develop and analyze a simplified asymptotic system, valid near special resonant flow states to be described below. We then show that this asymptotic model can reproduce in qualitative fashion many of the scenarios of DDT documented for the multi-phase flow system through the use of large scale computing [1, 2, 3]. This paper reviews some of the results I obtained in collaboration with A. Majda, M. Baer, and J. Hunter [6, 7, 8], and it is organized as follows. In Section 2 we present briefly the characteristics analysis for the reactive multi-phase flow equations of Baer and Nunziato and exhibit the singular resonant points. In Section 3 we discuss the asymptotic model for hot spot formation valid near resonant points. Finally, in Section 4 we analyze within the context of the asymptotic model, the nonlinear mechanisms for the development of hot spots.

## 2 Characteristics analysis of the multi-phase system and resonant points

The reactive multi-phase flow equations formulated by Baer and Nunziato are based on the continuum theory of mixtures, where both the solid and gas phases are assumed compressible and in thermodynamic non-equilibrium. An important difference between their formulation from previous systems for two-phase flows is in how the closure of the system is obtained. In addition to the usual saturation constraint and equations of state for each phase, the solid volume fraction is recognized as an independent degree of freedom, and its evolution is described by an equation consistent with the second law of thermodynamics [1].

More specifically, let the subscript  $a$  denote either the solid ( $a=s$ ) or the gas ( $a=g$ ); the Baer-Nunziato system consists of the following 7 evolution equations [1, 2] :

**Conservation of mass:**

$$\frac{\partial}{\partial t}(\phi_a \rho_a) + \frac{\partial}{\partial x}(\phi_a \rho_a v_a) = c_a^\dagger \quad (1)$$

**Conservation of momentum:**

$$\frac{\partial}{\partial t}(\phi_a \rho_a v_a) + \frac{\partial}{\partial x}(\phi_a (\rho_a v_a^2 + p_a)) - p_g \frac{\partial \phi_a}{\partial x} = m_a^\dagger \quad (2)$$

**Conservation of energy:**

$$\frac{\partial}{\partial t}(\phi_a \rho_a E_a) + \frac{\partial}{\partial x}(\phi_a(\rho_a E_a + p_a)v_a) - p_s \frac{\partial \phi_a}{\partial x} v_s = e_a^\dagger \quad (3)$$

**Compaction equation for the solid volume fraction:**

$$\frac{\partial \phi_s}{\partial t} + v_s \frac{\partial \phi_s}{\partial x} = \frac{c_s^\dagger}{\rho_s} + f_c \quad (4)$$

**Saturation constraint:**

$$\phi_s + \phi_g = 1 \quad (5)$$

**Equations of state for each phase:**

$$e_a = e_a(\rho_a, p_a). \quad (6)$$

Associated to the  $a$ -phase is the velocity  $v_a$ , material density  $\rho_a$ , pressure  $p_a$ , temperature  $T_a$ , internal energy  $e_a$  and volume fraction  $\phi_a$ . The total energy is given by  $E_a = e_a + v_a^2/2$ .

For simplicity we assume that both phases are described by equations of state of ideal gas type

$$\begin{aligned} e_a &= \frac{p_a}{\rho_a \Gamma_a} \\ T_a &= \frac{e_a}{c_v^a}, \end{aligned} \quad (7)$$

with  $c_v^a$ ,  $\Gamma_a$  constants.  $\Gamma_a$  is the Grüneisen coefficient and it is related to the  $\gamma$ -gas constant  $\gamma_a$  by  $\Gamma_a = \gamma_a - 1$ . Representative values for the gas and the solid are  $\Gamma_g = 0.2$  and  $\Gamma_s = 3$ .

Next we discuss the phase interaction terms  $c_s^\dagger$ ,  $m_s^\dagger$  and  $e_s^\dagger$ . First of all, since mass, momentum and energy for the mixture have to be conserved, the phase interaction terms must satisfy the constraints

$$\begin{aligned} c_g^\dagger &= -c_s^\dagger \\ m_g^\dagger &= -m_s^\dagger \\ e_g^\dagger &= -e_s^\dagger. \end{aligned} \quad (8)$$

The reaction term  $c_s^\dagger$  gives the rate of depletion of the solid due to surface burning and for simplicity we assume a one-step forward reaction with Arrhenius kinetics

$$c_s^\dagger = -K \rho_s \exp \left( A \left( \frac{1}{T_i} - \frac{1}{T} \right) \right), \quad (9)$$

here  $A$  is the activation energy,  $T_i$  is a reference temperature, and  $K$  is the pre-exponential factor. All are assumed to be constants.  $T$  is the mixture temperature and it is given by  $T = \phi_s T_s + \phi_g T_g$ .

The second term  $f_c$  in the compaction equation represents changes in volume fraction due to pressure differences, and it is given by

$$f_c = \frac{\phi_s \phi_g}{\mu_c} (p_s - p_g - \beta_s(\phi_s)), \quad (10)$$

where  $\mu_c$  is a compaction viscosity coefficient and  $\beta_s(\phi_s)$  represents intergranular stress due to grain contact.

The phase interaction terms  $m_s^\dagger$  and  $e_s^\dagger$  for the exchange of momentum and energy are given by

$$\begin{aligned} m_s^\dagger &= - \left( \delta + \frac{c_s^\dagger}{2} \right) (v_s - v_g) + c_s^\dagger v_s, \\ e_s^\dagger &= - \left( \delta + \frac{c_s^\dagger}{2} \right) (v_s - v_g) v_s - (p_s - \beta_s) f_c - h(T_s - T_g) + E_s c_s^\dagger. \end{aligned} \quad (11)$$

The terms in Eq. 11 involving  $c_s^\dagger$  and  $f_c$  represent exchange in momentum and energy due to chemical reaction and compaction respectively.  $\delta(v_s - v_g)$  represents changes in momentum due to drag, and  $h(T_s - T_g)$  represents changes in energy due to convective heat transfer between the gas and the solid.

Next we discuss the characteristics for the reactive multi-phase system. All the algebraic details can be found in [6, 7]. With  $u = (\rho_s, v_s, p_s, \phi_s, \rho_g, v_g, p_g)^T$ , Eqns. 1-4 can be written in matrix form as

$$A_0(u) \frac{\partial u}{\partial t} + A_1(u) \frac{\partial u}{\partial x} = S(u), \quad (12)$$

where  $A_0(u)$  and  $A_1(u)$  have the block structure

$$\begin{aligned} A_0(u) &= \left( \begin{array}{c|c|c} \phi_s \nabla f_0(u_s) & f_0(u_s) & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -f_0(u_g) & \phi_g \nabla f_0(u_g) \end{array} \right) \\ A_1(u) &= \left( \begin{array}{c|c|c} \phi_s \nabla f_1(u_s) & f_1(u_s) + e(u) & 0 \\ \hline 0 & v_s & 0 \\ \hline 0 & -f_1(u_g) - e(u) & \phi_g \nabla f_1(u_g) \end{array} \right), \end{aligned} \quad (13)$$

with  $f_0(u_a)$ ,  $f_1(u_a)$  and  $e(u)$  given by ( $a = s, g$ ):

$$\begin{aligned} f_0(u_a) &= (\rho_a, \rho_a v_a, \rho_a E_a)^T, a = s, g \\ f_1(u_a) &= (\rho_a v_a, \rho_a v_a^2 + p_a, (\rho_a E_a + p_a) v_a)^T, a = s, g \\ e(u) &= (0, -p_g, -p_g v_s)^T. \end{aligned} \quad (14)$$

The characteristic analysis of the system is determined by the study of the eigenvalues  $\lambda_j(u)$  and their associated right eigenvectors  $r_j(u)$ ,  $j = 1, \dots, 7$ , of the matrix equation

$$(-\lambda_j(u)A_0(u) + A_1(u))r_j(u) = 0, \quad (15)$$

and they are given by:

**Eigenvalues**

$$\begin{aligned} \lambda_1 &= v_g + c_g, & \lambda_2 &= v_s, & \lambda_3 &= v_s, & \lambda_4 &= v_g - c_g, \\ \lambda_5 &= v_g, & \lambda_6 &= v_s + c_s, & \lambda_7 &= v_s - c_s \end{aligned} \quad (16)$$

**Right Eigenvectors**

$$\begin{aligned} r_1 &= \left( 0, 0, 0, 0, 1, \frac{c_g}{\rho_g}, c_g^2 \right)^T \\ r_2 &= \left( 0, 0, -\frac{\phi_g(p_s - p_g)}{\phi_s \rho_g c_g^2} [(v_s - v_g)^2 - c_g^2], \frac{\phi_g}{\rho_g c_g^2} [(v_s - v_g)^2 - c_g^2], \right. \\ &\quad \left. \frac{(v_s - v_g)^2}{c_g^2}, \frac{v_s - v_g}{\rho_g}, (v_s - v_g)^2 \right)^T \\ r_3 &= (1, 0, 0, 0, 0, 0, 0)^T \\ r_4 &= \left( 0, 0, 0, 0, 1, -\frac{c_g}{\rho_g}, c_g^2 \right)^T \\ r_5 &= (0, 0, 0, 0, 1, 0, 0)^T \\ r_6 &= \left( 1, \frac{c_s}{\rho_s}, c_s^2, 0, 0, 0, 0 \right)^T \\ r_7 &= \left( 1, -\frac{c_s}{\rho_s}, c_s^2, 0, 0, 0, 0 \right)^T. \end{aligned} \quad (17)$$

Clearly the system is hyperbolic, with the wave speeds corresponding to the standard acoustic and particle speeds for each one of the phases. This fact is important both from the physical and computational grounds. In fact, several of the older models proposed for multi-phase flows exhibited complex characteristics and elliptic regions at low speed regimes [5, 13]. This triggers catastrophic instabilities of Hadamard type at the initial deflagration stage and invalidates the results of numerical calculations. The right eigenvectors of the system are also the standard acoustic and entropy modes for each phase, with the exception of the eigenvector  $r_2$  does not have an analog in one-phase flow and corresponds to compaction waves. Although the system is hyperbolic it is not strictly hyperbolic because  $v_s$  is a double eigenvalue, and also because the wave speeds associated with each phase change independently of each other, so that the eigenvalues of the solid phase can coincide with any of those for the gas phase. However, the system is totally hyperbolic: it has a complete family of associated right eigenvectors, except at special singular points. These singular resonant points are realized when one of the acoustic signals for the gas equals the solid particle speed:  $v_s = v_g \pm c_g$ . In

this case the compaction mode degenerates into one of the gas-acoustic modes:

$$\begin{aligned}\lim_{v_s \rightarrow v_g + c_g} r_2(u) &= r_1(u) \\ \lim_{v_s \rightarrow v_g - c_g} r_2(u) &= r_4(u).\end{aligned}\tag{18}$$

The singular resonant points  $v_g \pm c_g = v_s$  are also known as choked flow points [6, 7] because at the pore level they can be interpreted as choking of the gas flow through the pore space between grains. For definiteness consider the first case where  $r_1$  and  $r_2$  are aligned. The other case is analogous. In this case  $v_s$  is a triple eigenvalue with only two eigenvectors  $r_1$  and  $r_3$ . In fact, there is now a  $2 \times 2$  Jordan block structure coupling the acoustic eigenvector  $r_1$  with a generalized compaction eigenvector which we still denote by  $r_2$ :

$$\begin{aligned}(-\lambda A_0(u) + A_1(u))r_1 &= 0 \\ (-\lambda A_0(u) + A_1(u))r_2 &= -A_0(u)r_1 \\ (-\lambda A_0(u) + A_1(u))r_3 &= 0,\end{aligned}\tag{19}$$

where  $r_2$  is given explicitly by

$$r_2 = \left(0, 0, \frac{2\phi_g}{\rho_g c_g} \frac{p_s - p_g}{\phi_s}, -\frac{2\phi_g}{\rho_g c_g}, 0, \frac{1}{\rho_g}, 0\right)^T.\tag{20}$$

The algebraic structure displayed in Eq. 19 already puts in evidence the resonant character of the choked flow points through the coupling of the gas-acoustic and the compaction eigenvectors. However, the solid entropy mode remains decoupled from the other two. Therefore near choked flow conditions there is strong nonlinear interaction due to the coupling the gas-acoustics and compaction modes; it is near these resonant points that the simplified asymptotic model will be formulated in the next section.

### 3 The asymptotic model near a resonant state

The full derivation of the asymptotic model near resonant points is given in [7]. The interested reader can study the details of the lengthy derivation there. Here we just want to comment on the physical and mathematical basis of the derivation and connections of the asymptotic model and the continuum reactive multi-phase system. The purpose of the asymptotic model is to describe the nonlinear resonant interaction of the gas-acoustics and compaction modes at the convective stage of DDT, with a fast moving wave satisfying near choked flow and ignition temperature conditions. The physical mechanisms singled out are the resonance of the gas-acoustic and compaction modes, the burning of the solid explosive, and the convective heat transfer from the hot gas products to the solid combustible. The mathematical derivation of the model is asymptotic in nature and combines the methods of nonlinear geometrical optics with large activation energy asymptotics [9, 10, 11, 12]. Therefore we assume: 1. a resonant background state  $u_0$  with  $v_s^0 = v_g^0 + c_g^0$ . 2. near ignition temperature conditions for the background state. 3. Large activation energy for the reac-

tion, and 4. weak chemistry for the reaction. This last requirement is due to the small amplitude nature of the asymptotic approximation. The reactive multi-phase equations can be cast in nondimensional form as follows (see [7] for the details of the scaling procedure):

$$A_0(u) \frac{\partial u}{\partial t} + A_1(u) \frac{\partial u}{\partial x} = \epsilon S(u, u_0, \epsilon). \quad (21)$$

The relevant nondimensional parameter  $\epsilon$  is proportional to the inverse of the activation energy, so that  $\epsilon \ll 1$ . The nondimensional matrices  $A_0(u)$  and  $A_1(u)$  have the same form as in Eqs. 12-13. In the nondimensional source term  $S(u, u_0, \epsilon)$  the dominant effect is due by heat convection, followed by chemical reaction and drag:

$$\epsilon S(u, u_0, \epsilon) = \epsilon S^1(u) + \epsilon^2 [\mathcal{W}(u, u_0, \epsilon) S^2(u) + \tilde{S}^2(u)] + \epsilon^3 S^3(u), \quad (22)$$

where

$$\begin{aligned} S^1(u) &= \tilde{h}(T_s - T_g) \cdot (0, 0, -1, 0, 0, 0, 1)^T \\ S^2(u) &= \left( 1, \frac{v_s + v_g}{2}, \epsilon_s + \frac{v_s v_g}{2}, \frac{1}{\rho_s}, -1, -\frac{v_s + v_g}{2}, -\epsilon_s - \frac{v_s v_g}{2} \right)^T \\ \mathcal{W}(u, u_0, \epsilon) &= \tilde{k}_r \exp \left( \frac{1}{\epsilon} \left( \frac{1}{T_0} - \frac{1}{T} \right) \right), T_0 = 1 \\ \tilde{S}^2(u) &= \tilde{\delta}(v_s - v_g) \cdot (0, -1, -v_s, 0, 0, 1, v_s)^T \\ S^3(u) &= \tilde{f}_c \cdot (0, 0, -p_s + \tilde{\beta}_s, 1, 0, 0, p_s - \tilde{\beta}_s)^T \\ \tilde{f}_c(u) &= \tilde{k}_c \phi_s \phi_g (p_s - p_g - \tilde{\beta}_s). \end{aligned} \quad (23)$$

In the nonlinear geometrical optics approximation we assume small amplitude high frequency waves around the background resonant point  $u_0$  with resonant speed  $\lambda = v_s^0 = v_g^0 + c_g^0$ . With the fast variable  $\theta = (x - \lambda t)/\epsilon$  the *ansatz* of nonlinear geometrical optics becomes

$$u'(\theta, t) = u_0 + \epsilon(\bar{u}_1(t) + \sigma_1(\theta, t)r_1) + \epsilon^2(\sigma_2(\theta, t)r_2 + \dots) + O(\epsilon^3). \quad (24)$$

Here  $\bar{u}_1(t)$  represents the mean-field effects due to convective heat transfer,  $\sigma_1(\theta, t)$  is the gas-acoustic perturbation amplitude, and  $\sigma_2(\theta, t)$  is the compaction perturbation amplitude. The resonant character of the approximation is already evident at the formal asymptotic level, where the order  $\epsilon^2$  compaction term resonates with the order  $\epsilon$  gas-acoustic term. We also omitted for simplicity the contributions of the solid entropy mode in the asymptotic approximation because from the algebraic structure of the system at resonant points given in Eq. 19, it is apparent that the coupling of the solid entropy mode with the other two is going to weak, and only through the source term. Nevertheless, the solid entropy mode is considered in the derivation given in [7].

The initial data is a perturbation around the resonant background state  $u_0$

$$u'(x, 0) = u_0 + \epsilon \left( \bar{u}^1 + \sigma_1^0 \left( \frac{x}{\epsilon} \right) r_1 \right) + \epsilon^2 \sigma_2^0 \left( \frac{x}{\epsilon} \right) r_2. \quad (25)$$

where the initial mean field perturbation  $\bar{u}^1$  at order  $\epsilon$  is given by  $\bar{u}^1 = (\bar{p}_s^1, \bar{v}_s^1, \bar{p}_g^1, \bar{\phi}_s^1, \bar{p}_g^1, \bar{v}_g^1, \bar{p}_g^1)^T$ ,  $r_1$  is given by Eq. 17 and  $r_2$  by Eq. 20. Without going into the lengthy details of the

derivation given in [7], the resulting asymptotic system for the resonant interaction of the gas-acoustic and compaction mode is given by:

$$\begin{pmatrix} \rho_s^\epsilon \\ v_s^\epsilon \\ p_s^\epsilon \\ \phi_s^\epsilon \\ \rho_g^\epsilon \\ v_g^\epsilon \\ p_g^\epsilon \end{pmatrix} = \begin{pmatrix} \rho_s^0 + \epsilon \bar{p}_s^1 + O(\epsilon^2) \\ v_g^0 + c_g^0 + \epsilon \bar{v}_s^1 + O(\epsilon^2) \\ p_s^0 + \epsilon (\bar{p}_s^1 + t \Gamma_s \tilde{h}(T_g^0 - T_s^0)/\phi_s^0) + O(\epsilon^2) \\ \phi_s^0 + \epsilon \bar{\phi}_s^1 - \epsilon^2 2\phi_g^0 \sigma_2^0(\theta, t)/\rho_g^0 c_g^0 + O(\epsilon^3) \\ \rho_g^0 + \epsilon (\bar{\rho}_g^1 + \sigma_1(\theta, t)) + O(\epsilon^2) \\ v_g^0 + \epsilon (\bar{v}_g^1 + c_g^0 \sigma_1(\theta, t))/\rho_g^0 + O(\epsilon^2) \\ p_g^0 + \epsilon (\bar{p}_g^1 - t \Gamma_g \tilde{h}(T_g^0 - T_s^0)/\phi_g^0 + (c_g^0)^2 \sigma_1(\theta, t)) + O(\epsilon^2) \end{pmatrix}, \quad (26)$$

where the gas-acoustic amplitude  $\sigma_1(\theta, t)$  and the compaction amplitude  $\sigma_2(\theta, t)$  satisfy the *simplified asymptotic equations*:

**Gas-acoustic mode:**

$$\frac{\partial \sigma_1}{\partial \theta} + \frac{\partial}{\partial \theta} \left( (a_1 - b_1 t) \sigma_1 + c_1 \frac{\sigma_1^2}{2} \right) = \frac{\partial \sigma_2}{\partial \theta} + \kappa_1 \frac{\partial^2 \sigma_1}{\partial \theta^2} \quad (27)$$

**Compaction mode:**

$$\frac{\partial \sigma_2}{\partial \theta} = R \exp(\alpha_1 \sigma_1 + \beta_1 + \beta_2 t), \quad (28)$$

Eqs. 27-28 have been written in a coordinate system that makes  $\sigma_2$  stationary. The coefficients in Eqs. 27-28 are given explicitly by

**Wave speed coefficients**

$$\begin{aligned} c_1 &= \frac{\Gamma_g + 2}{2} \frac{c_g^0}{\rho_g^0} > 0 \\ a_1 &= \bar{v}_g^1 - \bar{v}_s^1 + \frac{c_g^0}{2} \left( \frac{\bar{p}_g^1}{\rho_g^0} - \frac{\bar{p}_s^1}{\rho_g^0} \right) \\ b_1 &= \frac{c_g^0 \Gamma_g}{2 \phi_g^0 \rho_g^0} \tilde{h}(T_g^0 - T_s^0) \end{aligned} \quad (29)$$

### Source term coefficients

$$\begin{aligned}
 \alpha_1 &= \frac{\phi_s^0 p_g^0}{c_v^0 (\rho_g^0)^2} > 0 \\
 \beta_1 &= \phi_s^0 T_s^0 \left( \frac{\bar{p}_s^1}{p_s^0} - \frac{\bar{p}_s^1}{\rho_s^0} \right) + \phi_g^0 T_g^0 \left( \frac{\bar{p}_g^1}{p_g^0} - \frac{\bar{p}_g^1}{\rho_g^0} \right) - \bar{\phi}_s^1 (T_g^0 - T_s^0) \\
 \beta_2 &= \bar{h} (T_g^0 - T_s^0) \left( \Gamma_s \frac{T_s^0}{p_s^0} - \Gamma_g \frac{T_g^0}{p_g^0} \right) \\
 R &= \tilde{k}_r \frac{\rho_g^0 p_g^0 c_g^0}{2 \phi_g^0} > 0.
 \end{aligned} \tag{30}$$

From Eq. 26 it follows that an increase of the gas-acoustic amplitude  $\sigma_1$  corresponds to the increase of the physical variables  $\rho_g$  and  $p_g$ . Similarly, an increase of the compaction amplitude  $\sigma_2$  corresponds to a decrease of the volume fraction  $\phi_s$  of the solid, i.e. to burning of the solid. These will be relevant when studying the numerical solutions of the asymptotic system in the next section.

Of all the parameters in Eqs. 29-30 the most relevant in our analysis are  $b_1$  and  $\beta_2$ . Both parameters include convective heat transfer effects from the background state  $u_0$ .  $b_1$  controls the linear recession speed from the burning front in Eq. 27. If  $b_1 > 0$  then there is drift away from the burning front and the opposite occurs if  $b_1 < 0$ . It is clear from Eq. 29 that

$$b_1 > 0 \quad \text{iff} \quad T_g^0 > T_s^0, \tag{31}$$

that is, if at the background state the product gases preheat the solid by heat convection.  $b_1$  also is relevant in studying the resonant behavior that the asymptotic system inherits from the multi-phase equations. The measure of how near the flow is from resonance is given by the *relative Mach number*  $M$  defined by  $M = (v_s - v_g)/c_g$ . In terms of the asymptotic approximation in Eq. 26 is given by

$$M = 1 + \epsilon(b_1 t - c_1 \sigma_1 - a_1)/c_g^0 + O(\epsilon^2). \tag{32}$$

Hence, if a burning front initially has  $M < 1$  but  $b_1 > 0$  we expect that it will accelerate and go through the resonant state  $M = 1$  in finite time. In the next section we will specify the initial data so that  $M < 1$ .

The other relevant parameter is  $\beta_2$ . From Eq. 28 it follows that the chemical reaction is accelerated if  $\beta_2 > 0$  and inhibited otherwise. Therefore  $\beta_2$  is an important parameter to consider in the creation of hot spots. From Eq. 30 it follows that

$$\beta_2 > 0 \quad \text{iff} \quad 1 < \frac{T_g^0}{T_s^0} < \frac{\Gamma_s}{\Gamma_g}, \tag{33}$$

that is, in order to enhance the reaction the gas needs to be hotter than the solid but the temperatures cannot be too disparate, with a bound of their ratio in Eq. 33 given by the Grüneisen coefficients. With  $\Gamma_s \approx 3.0$  and  $\Gamma_g \approx 0.2$  we have an upper bound in Eq. 33 of the order of 60.



## 4 Hot spot development in the asymptotic model

In this section we discuss the qualitative predictions that can be made with the asymptotic model. The discussion is based on the paper by Embid and Majda [8] where the interested reader is referred for a more complete discussion of the cases presented here as well as additional cases of interest. The reader can find also in [8] a discussion of the numerical method used to solve the asymptotic system as well as interesting analytical solutions of the equations for the non-reactive case. For the experimental set-up we consider the situation where a fast moving burning front is initially near resonance. More specifically, we assume at the start a fairly stiff situation where there is no gas-acoustic disturbances produced, so that  $\sigma(\theta, 0) = 0$ , and the compaction mode is a monotonic decreasing profile as depicted in Fig. 1, so that there is more compaction on the right. The initial data is chosen so that the wave is subsonic ( $M < 1$ ) but near resonance. From Eqs. 29 and 32 it is clear that this can be achieved with  $u_0$  a resonant state and by adjusting the initial mean field correction  $\bar{u}^1$  so that  $a_1 > 0$ . With this prescription of the initial data we consider three cases of interest. In case 1 there is no chemical reaction. Case 2 has chemical reaction enhanced with a choice of  $\beta_2 > 0$ . Finally case 3 has inhibited reaction with  $\beta_2 < 0$ . The numerical solution of the Eqs. 27-28 was done with operator splitting and a higher order Godunov scheme for the nonlinear wave equation. For the details the reader is referred to [8].

**Case 1:**  $R = 0$ ,  $b_1 = 5$  (no reaction). In this case it is clear from Eq. 28 that the compaction amplitude  $\sigma_2$  does not change in time and it remains the monotonic decreasing profile in Fig. 1. On the other hand, the inhomogeneities in  $\sigma_2$  act as a source term for the gas-acoustic amplitude in Eq. 26. Overlays of the gas-acoustic amplitude in time are depicted in Fig. 2. From an initial uniform zero state merges a wave that grows initially in amplitude. Because  $b_1 > 0$  the wave moves to the left of the burning front. From Eq. 32 the wave also becomes resonant in finite time and develops a shock wave. Afterwards the wave amplitude decays in time.

**Case 2:**  $R = 1$ ,  $b_1 = 5$ ,  $\beta_2 = 2$  (enhanced reaction). Overlays in time of the solution for the amplitudes  $\sigma_1$  and  $\sigma_2$  are depicted in Figs. 3(a) and 3(b). At the initial stages from  $t = 0.0$  to  $t = 2.0$  the behavior of  $\sigma_1$  is similar to case 1 with a shock wave moving to the left of the burning front. At this stage  $\sigma_2$  shows essentially uniform burning throughout but with the presence of small disturbances produced by the left moving gas-acoustic shock wave. However, by time  $t = 2.5$  the feedback mechanism built in the asymptotic system produces amplification of the gas-acoustic mode and the formation of a region of relative compaction in  $\sigma_2$ . The resonant feedback mechanism enhances the growth of this gas-acoustic hot-spot and at the same time creates a zone of enhanced burning of the solid ahead of the region of relative compaction, represented by the spike in  $\sigma_2$  at times  $t = 2.8$  and  $2.81$ . At this stage the resonant feedback between both modes is very strong and induces the very rapid growth of the hot spot. Shortly after, by time  $t = 2.816$  and at location  $x = -14.9$ , the maximum measured amplitude for  $\sigma_1$  is about  $3.3 \times 10^3$  while the amplitude for  $\sigma_2$  yields values of (machine) infinity. Clearly, at this point we are beyond of the regime of validity of the asymptotic approximation and possibly other physical effects not incorporated in the model, such as compression of the solid may be relevant. We point out that the behavior described in this case is qualitatively similar to the situation documented in [1] for the combustion of a column of HMX with 95 % initial density. The numerical solution of the reactive-multiphase

equations shows the birth and growth of a hot spot during the convective stage of DDT (from 10-20  $\mu\text{sec.}$  and this hot spot moves away from the compaction front at a linear speed, in accordance with the predictions of the asymptotic model. We also remark that in [8] we discuss other parameter choices where the hot spot grows within the compaction burning front. This situation has also been observed in the multi-phase flow equations [1, 2, 3].

**Case 3:**  $R = 1$ ,  $b_1 = 5$ ,  $\beta_2 = -0.1$  (inhibited reaction). Overlays in time of the solution for the amplitudes  $\sigma_1$  and  $\sigma_2$  are depicted in Figs. 4(a) and 4(b). In this case we notice the strong inhibiting effect that a negative value of  $\beta_2$  has on the resonance mechanism and the creation of hot spots. In this case we notice that the gas-acoustic solution depicted in Fig. 4(a) does not develop a hot spot. In fact, the solution is almost identical to the nonreactive case depicted in Fig. 2. Also revealing is to observe the behavior of the compaction amplitude in Fig. 4(b). Here the compaction wave goes through essentially a very slow and uniform burning (compare with Fig. 3(b)), with the development at the initial stages of small disturbances that cannot amplify and rapidly die out. Situations qualitatively similar to the one described here are observed in stable fast deflagration waves that do not transit to detonation [3].

## 5 Conclusions

We have studied a state of the art continuum mixture theory system of equations for reactive multi-phase flow utilized in the study of DDT. We identified interesting singular states for the flow and using the methods of nonlinear geometrical optics and high activation energy asymptotics, derived a simplified asymptotic model to study the formation and growth of hot spots. We showed how by choosing different values of selected parameters one can reproduce in qualitative fashion different scenarios for DDT documented in the literature. By design the model applies only on the small amplitude regime and provides no insight on the later stages of DDT process. On the other hand, the model provides interesting parameter regimes with physical significance that indicate whether or not hot spots will develop and the relevant physical mechanisms responsible for it.

## References

- [1] M. Baer, and J. Nunziato, *A two-phase mixture theory for the deflagration-to-detonation transition (DDT) in reactive granular materials*, Int.J. of Multiphase Flow, 12 (1986), pp. 861-889.
- [2] M. Baer, R. Gross, J. Nunziato, and E. Igel, *An experimental and theoretical study of deflagration-to-detonation transition (DDT) in the granular explosive CP*, Combustion and Flame, 65 (1986), pp. 15-30.
- [3] M. Baer, J. Nunziato, and P. Embid, *Deflagration-to-detonation transition (DDT) in reactive granular materials*, Progress in Astronautics and Aeronautics, 135 (1991), pp. 481-512.

- [4] R. Berneker, *The deflagration-to-detonation transition of high energy propellants - a review*, AIAA Journal 24 (1986), pp. 82-91.
- [5] P. Embid, and M. Baer, *Modeling two-phase flow of reactive granular materials*, IMA Volumes Math. Applications, 29 (1991) 58-67.
- [6] P. Embid, and M. Baer, *Mathematical analysis of a two-phase model for reactive granular flow*, Continuum Mechanics and Thermodynamics, 4 (1992), pp. 279-312.
- [7] P. Embid, J. Hunter, and A. Majda, *Simplified asymptotic equations for the transition to detonation in reactive granular materials*, SIAM J. Applied Math., 52(5), pp. 1199-1237.
- [8] P. Embid, and A. Majda, *An asymptotic Theory for hot spot formation and transition for reactive granular materials*, Combustion and Flame, 89 (1992), pp. 17-36.
- [9] J. Hunter, and J. Keller, *Weakly nonlinear high frequency waves*, Comm. Pure Appl. Math., 12 (1983), pp. 543-569.
- [10] A. Majda, *High Mach number combustion*, Reacting Flows: Combustion and Chemical Reactors, AMS Lectures in Applied Mathematics, 24 (1986), pp. 109-184.
- [11] A. Majda, and R. Rosales, *Nonlinear mean field-high frequency wave interaction in the induction zone*, SIAM J. Appl. Math., 47 (1987), pp. 1017-1039.
- [12] R. Rosales, and A. Majda, *Weakly nonlinear detonation waves*, SIAM J. Appl. Math., 43 (1983), pp. 1086-1118.
- [13] H. Stewart, and B. Wendroff, *Two-phase flow: models and methods*, J. Comp. Phys., 56 (1984), pp. 363-409.

## Figure Captions

Fig. 1 Initial burning wave profile for the compaction mode  $\sigma_2$

Fig. 2 Overlays of the gas acoustic amplitude  $\sigma_1$  for the nonreactive wave with heat convection. Case  $b_1 = 5$ . Snapshots recorded from  $t = 0.0$  to  $t = 10.0$  at time intervals  $\Delta t = 1.0$ .

Fig. 3 Overlays for a reactive wave. Case  $b_1 = 5$ ,  $\beta_2 = 2$ . (a) Amplitude  $\sigma_1$ . (b) Amplitude  $\sigma_2$ . Snapshots recorded at times  $t = 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 2.6, 2.7, 2.8$ , and  $2.81$ .

Fig. 4 Overlays for a reactive wave. Case  $b_1 = 5$ ,  $\beta_2 = -0.1$ . (a) Amplitude  $\sigma_1$ . (b) Amplitude  $\sigma_2$ . Snapshots recorded from  $t = 0.0$  to  $t = 10.0$  at time intervals  $\Delta t = 1.0$ .

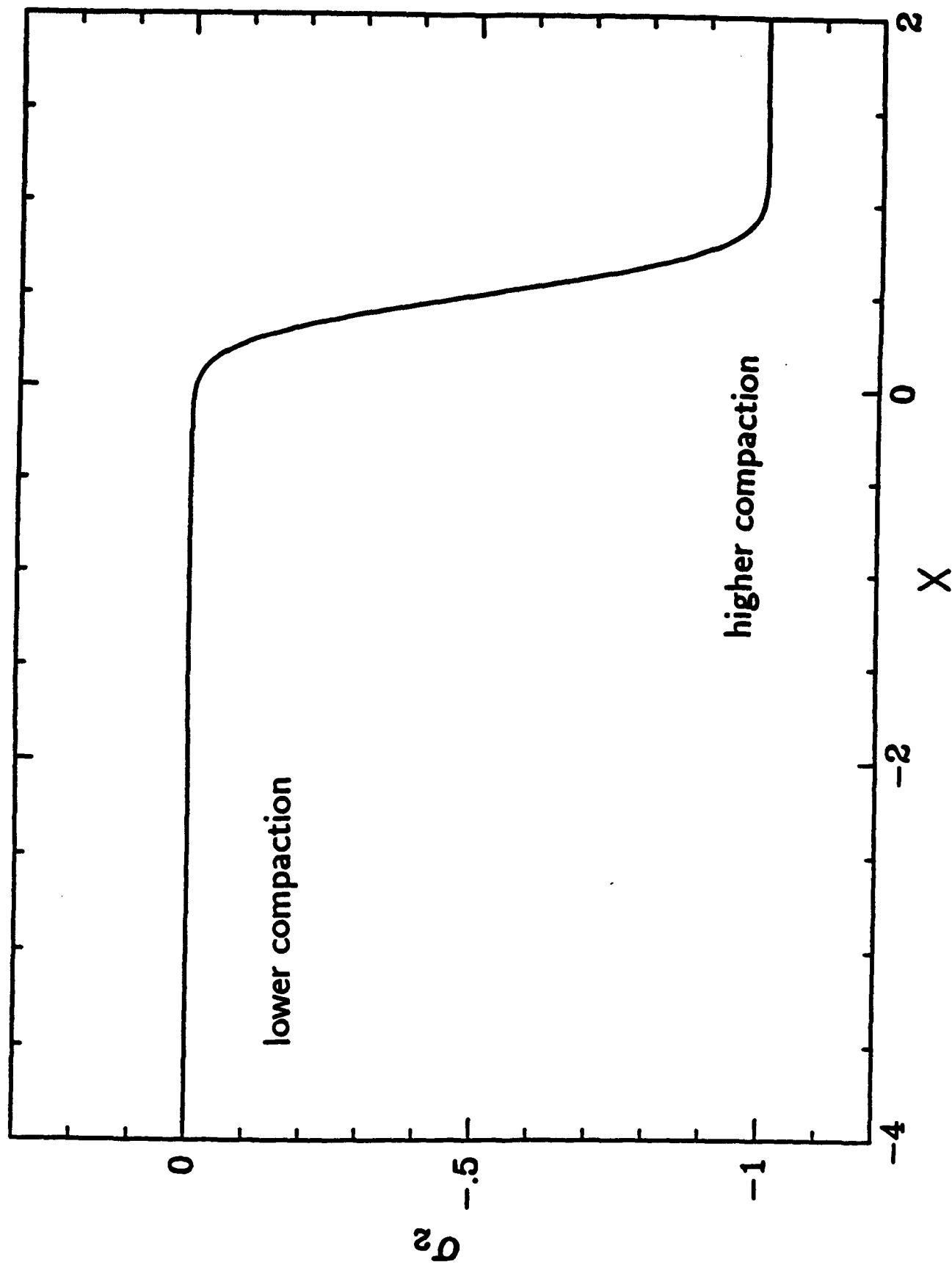


Fig. 1

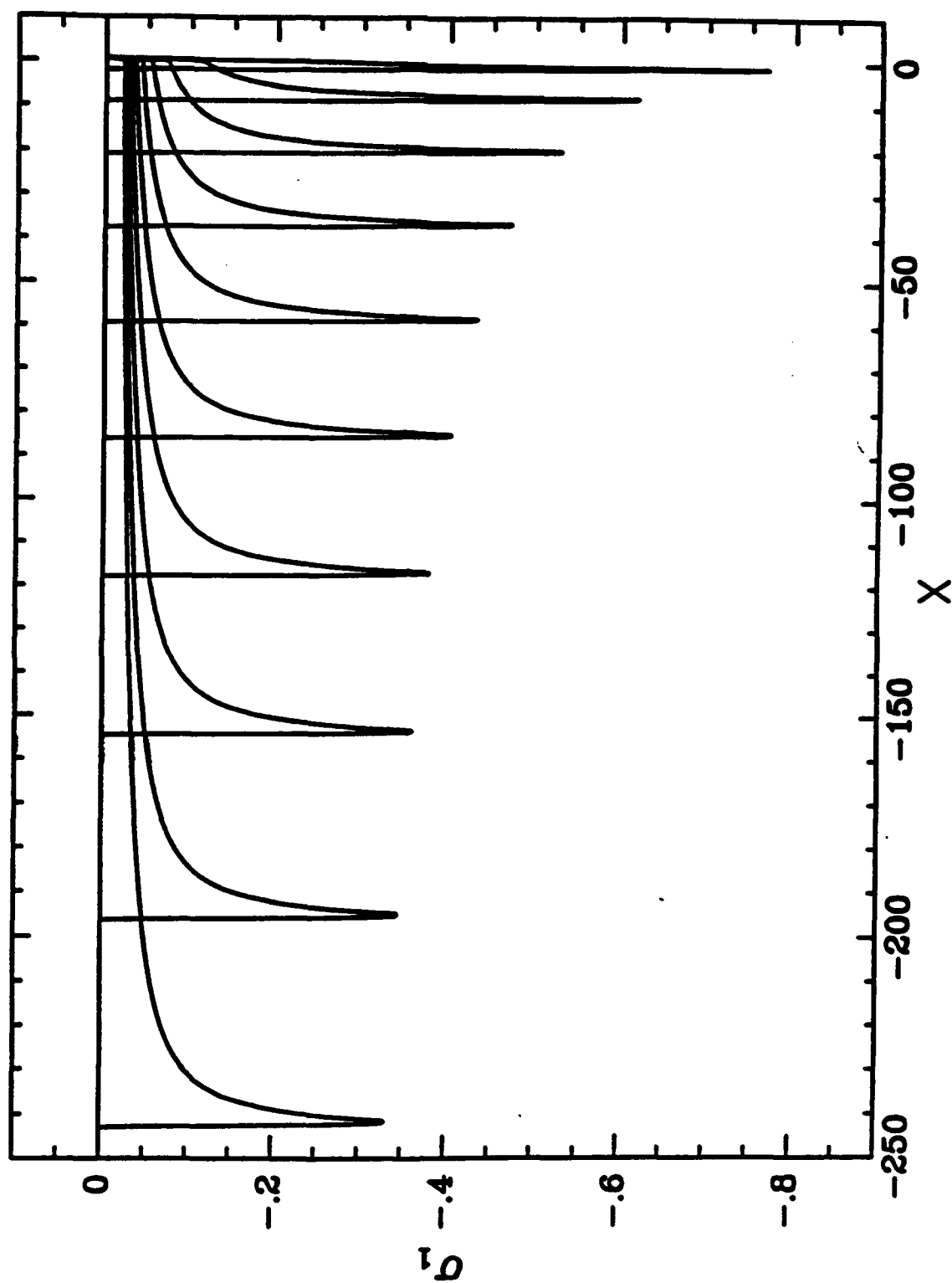


Fig. 2

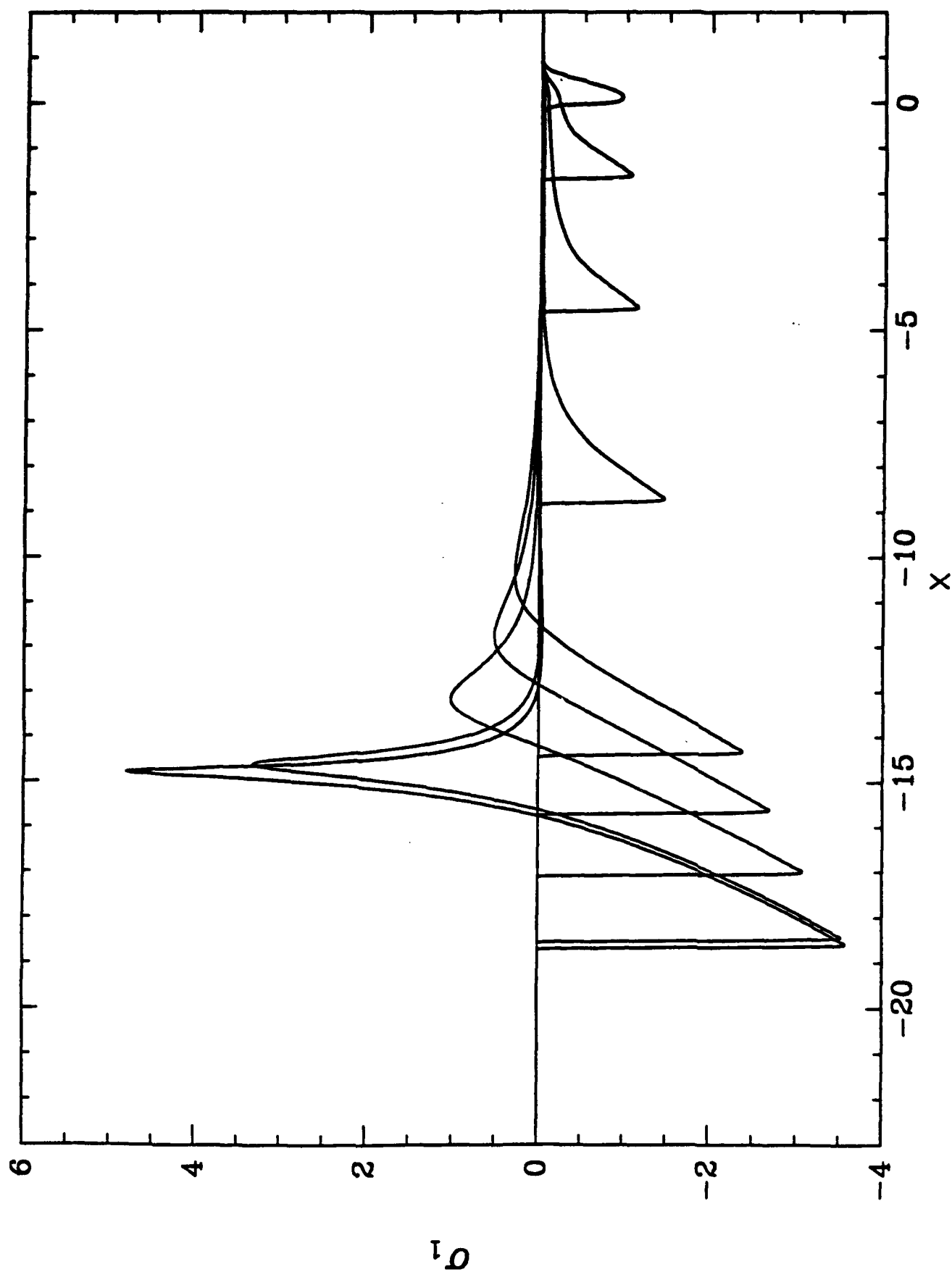


Fig. 3(a)

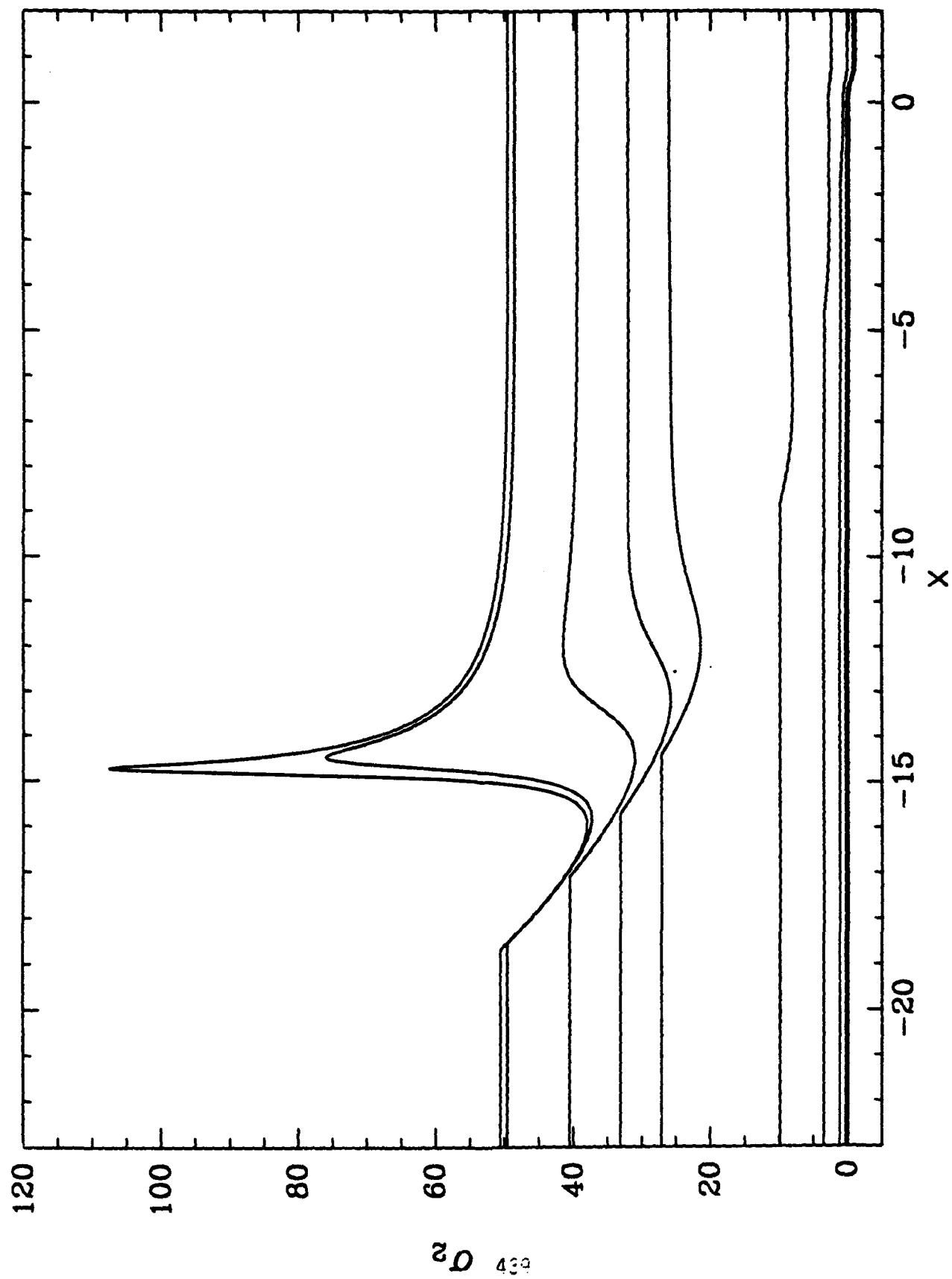


Fig. 3(b)



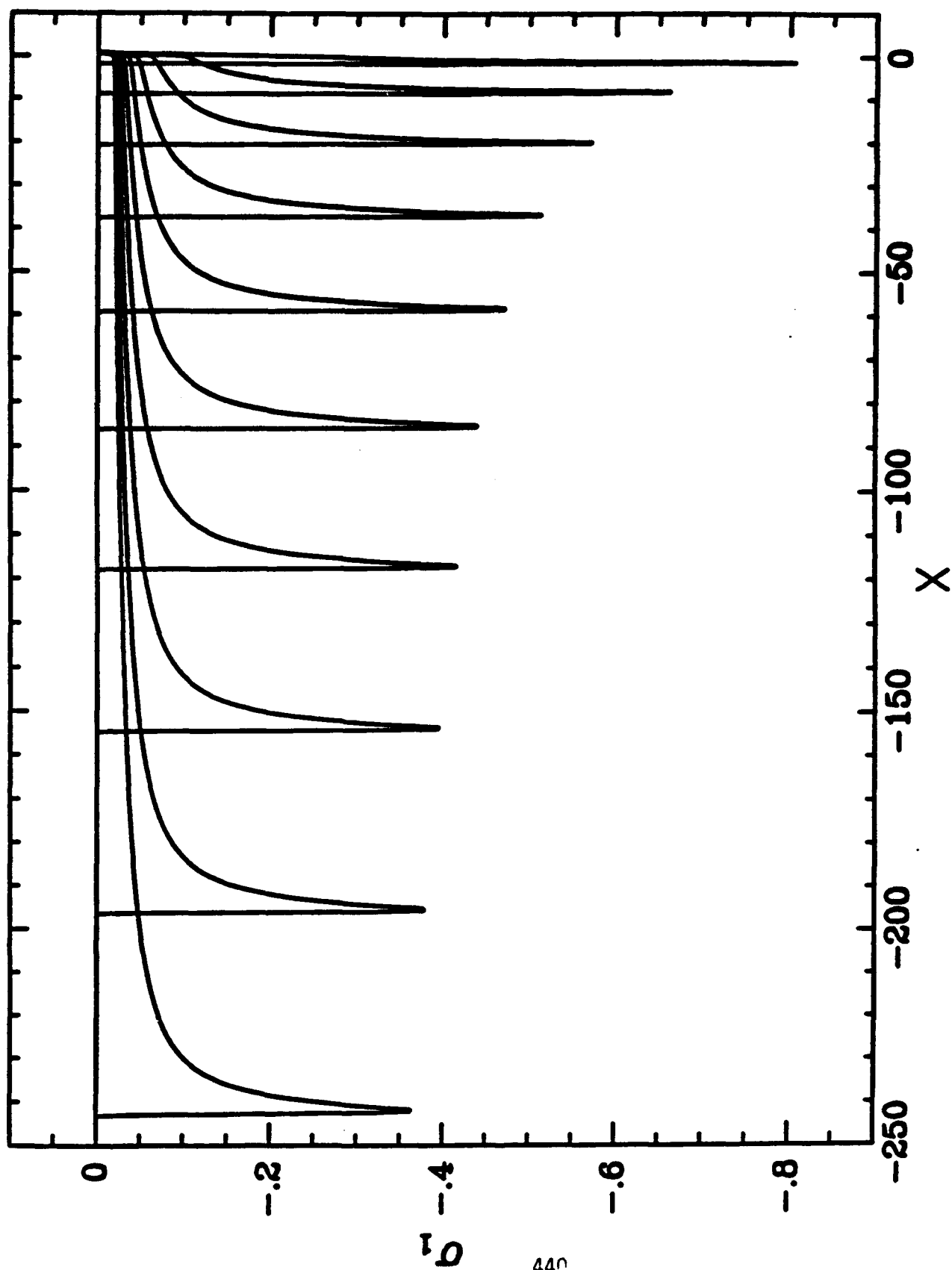


Fig. 4(a)

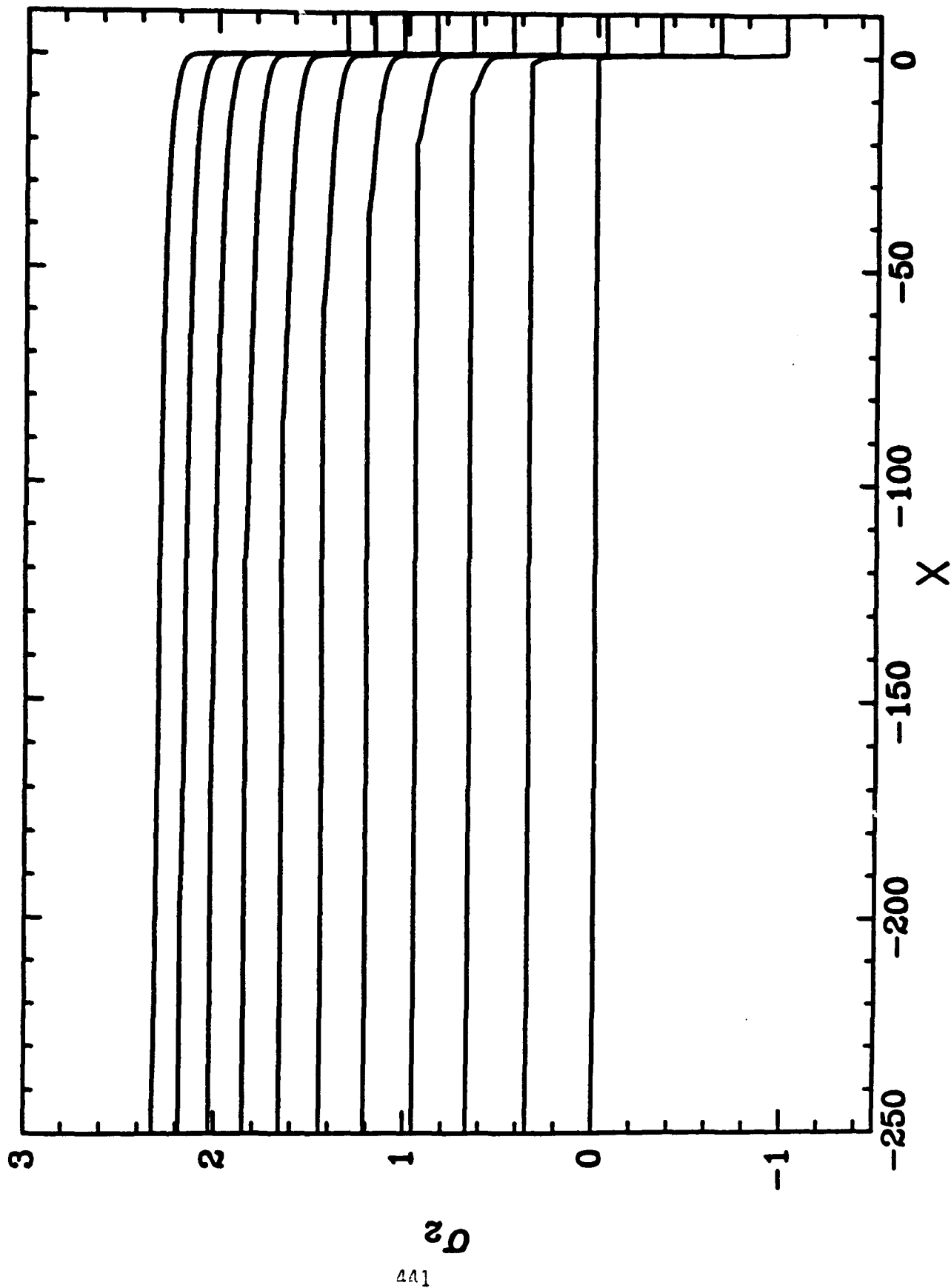


Fig. 4(b)

## THERMIONIC EMISSION FROM HIGH- $T_c$ SUPERCONDUCTORS

Richard A. Weiss

U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** The superconducting state of a high- $T_c$  material with  $T < T_c$  is described by a coherent spacetime state in which the electrons of a Cooper pair rotate coherently or incoherently in the azimuthal angle space and time variables, and move coherently in the radial space and time variables. In this sense the electrons of a Cooper pair are localized in the radial space and time coordinates. The normal state of a high- $T_c$  material with  $T > T_c$  occurs when the electrons in a Cooper pair are in a partially coherent radial spacetime state. The equilibrium equations for a pair of Cooper electrons are formulated for the coherent radial spacetime conditions of the superconducting state, and it is found that for a weak attractive Coulomb pairing force the rotation is non-Keplerian in the sense that the angular frequency of rotation is independent of the separation distance of the two electrons of the pair. Applications to the theory of thermionic emission from high- $T_c$  materials are considered, and the thermionic emission current is calculated for the normal and superconducting states. The thermionic emission current for the normal state of a high- $T_c$  material is found to be given by a modified Richardson-Dushman equation with  $T^2$  still as the leading temperature dependent term, while the thermoemission from the superconducting state is given by a totally different expression which does not have the leading quadratic temperature term but instead can have leading temperature dependent terms of the form  $T^0$ ,  $T^{1/2}$ ,  $T$  or  $T^{3/2}$ .

**1. INTRODUCTION.** The discovery of high- $T_c$  superconductors brings the promise of the development of a new technology in communications, power generation and transmission, and mass transportation.<sup>1-5</sup> High- $T_c$  superconductors for instance will lead to stronger electromagnets for research and industry. The prospective uses are limitless if the high- $T_c$  materials can be found that have important material properties such as ductility, malleability, strength and the ability to maintain the superconducting state in the presence of the magnetic field generated by the current flowing in the high- $T_c$  superconductor. The materials have to be designed to exhibit these useful engineering properties in addition to having the property of high- $T_c$  superconductivity. Trial and error design of such substances is of value, but considering the magnitude of the number of inorganic and organic compounds that need to be investigated it is clear that the trial and error approach should be supported by an analytical method for the design of high- $T_c$  compounds having engineering value.

New material design requires an understanding of the chemical and physical processes that occur in a material to produce a high- $T_c$  superconducting state. Unfortunately the underlying mechanism of high- $T_c$  superconductivity is not completely understood for the layered oxide, organic, or heavy fermion type of superconductors. This paper develops a theory of the thermionic emission from high- $T_c$  materials which can possibly lead to the development of an analytical understanding of the properties/structure relationship for these materials. Specifically, the paper investigates the motion of electrons in Cooper pairs that form in high- $T_c$  superconductors, and calculates the various forms of the therm-

ionic emission current from high- $T_c$  materials. The motion of electrons in high- $T_c$  materials are subject to the broken symmetry nature of spacetime in these materials, and the electrical properties of a high- $T_c$  material will depend on the type of broken spacetime symmetry that occurs in the Cooper electron pairs. The outline of this paper is as follows: Section 2 considers the non-Keplerian motion of the electrons in a Cooper pair in the superconducting state of a high- $T_c$  material, and Section 3 develops a theory of the thermionic emission from the normal and superconducting states of high- $T_c$  materials.

As in conventional BCS theory, high- $T_c$  superconductivity is generally interpreted as being due to a broken gauge symmetry associated with the formation of Cooper electron pairs in the ground state.<sup>6-9</sup> The electron-phonon interaction probably is responsible for the attractive force between two electrons that are located in the vicinity of lattice atoms.<sup>6-9</sup> However, these concepts have not been able to explain the large values of the measured normalized superconductivity energy gaps in high- $T_c$  materials which have the following range of values<sup>1-5,9-12</sup>

$$3 < 2\Delta/(kT_c) < 8 \quad (1)$$

The large values of the measured normalized superconductivity energy gaps for high- $T_c$  materials has been interpreted as being due to a strong binding of the electrons in a Cooper pair.<sup>1-5</sup> In fact the BCS theory of weakly interacting electrons in a Cooper pair predicts that the normalized superconductivity energy gap is given by

$$2\Delta/(kT_c) = 3.52 \quad (2)$$

where in this case the half-width gap  $\Delta$  refers to absolute zero temperature.

Recently a broken spacetime symmetry theory of high- $T_c$  materials was developed that assumes that for the superconducting state the space and time coordinates of the electrons in a Cooper pair change by a coherent rotation in an internal space, i.e. space and time are localized for the electrons in Cooper pairs.<sup>13,14</sup> This concept predicts that if the two electrons in a Cooper pair are weakly interacting the normalized superconductivity energy gap is nevertheless not given by the BCS value but instead has the value<sup>13,14</sup>

$$2\Delta/(kT_c) = 6/\pi(3.52)(1 - 4/\pi \theta_a)^{-1} \quad (3)$$

where  $\theta_a$  = internal phase angle of the relative acceleration of the two electrons in a Cooper pair. The factor  $6/\pi$  arises from the condition of superconductivity in a broken spacetime symmetry theory of Ohm's law and the assumption of free electrons (weak interaction) which require that the internal phase angle of coherent time has the value  $\theta_t = \pi/6$ .<sup>13</sup> The normal state of a high- $T_c$  material is interpreted to be a partially coherent spacetime state.<sup>13,14</sup> Laboratory measurements using a variety of experimental techniques on various types of high- $T_c$  materials has yielded values of the normalized superconductivity energy gap that are larger than the BCS value of 3.52 as indicated by equations (1) and (2).<sup>1-5,9-12,15,16</sup> This is true for the common oxide high- $T_c$  materials and for the alkali metal doped  $C_{60}$  fullerenes. These large values of the superconductivity energy gap are due to the factor  $6/\pi$  that is associated with the coherent radial motion of the electrons in the Cooper pairs of a weakly interacting high- $T_c$  material.<sup>13,14</sup>

Within the weakly coupled electron gas of a high- $T_c$  material the space and time coordinates are complex numbers in an internal space and can be written as<sup>17,18,20</sup>

$$\bar{t}_v = t_v \exp(j\theta_t^v) \quad (4)$$

$$\bar{v} = v \exp(j\theta_v) \quad (5)$$

where  $v = x, y, z$  for cartesian coordinates,  $v = r, \phi, z$  for cylindrical polar coordinates, and  $v = \rho, \phi, \psi$  for spherical polar coordinates. Strictly speaking a time coordinate is associated with each space coordinate as in equations (4) and (5), but for homogeneous spacetime it follows that

$$t_v = t \quad \theta_t^v = \theta_t \quad (6)$$

for all  $v$ . Essentially the broken symmetry of spacetime can be deduced from a relativistic trace equation for matter and energy.<sup>17-19</sup> The differentials of the time and space coordinates can be obtained from equations (4) and (5) to be<sup>17,18,20</sup>

$$d\bar{t}_v = \sec \beta_{tt}^v dt_v \exp(j\phi_t^v) = \csc \beta_{tt}^v t_v d\theta_t^v \exp(j\phi_t^v) \quad (7)$$

$$d\bar{v} = \sec \beta_{vv} dv \exp(j\phi_v) = \csc \beta_{vv} v d\theta_v \exp(j\phi_v) \quad (8)$$

where

$$\tan \beta_{tt}^v = t_v \partial \theta_t^v / \partial t_v \quad (9)$$

$$\tan \beta_{vv} = v \partial \theta_v / \partial v \quad (10)$$

$$\phi_t^v = \theta_t^v + \beta_{tt}^v \quad (11)$$

$$\phi_v = \theta_v + \beta_{vv} \quad (12)$$

Then the magnitudes of the differentials of space and time are given by<sup>17,18,20</sup>

$$|d\bar{t}_v| = \sec \beta_{tt}^v dt_v = \csc \beta_{tt}^v t_v d\theta_t^v \quad (13)$$

$$|d\bar{v}| = \sec \beta_{vv} dv = \csc \beta_{vv} v d\theta_v \quad (14)$$

and the measured values of the time and space coordinates are given by<sup>17,18,20</sup>

$$t_{vm} = t_v \cos \theta_t^v \quad (15)$$

$$v_m = v \cos \theta_v \quad (16)$$

where  $v = x, y, z$ ;  $r, \phi, z$  or  $\rho, \phi, \psi$ .

The velocity and momentum components of a particle moving in broken symmetry spacetime must also be written as complex numbers as follows<sup>17,18,20</sup>

$$\bar{v}_v = v_v \exp(j\theta_{vv}) = d\bar{v}/d\bar{t}_v \quad (17)$$

$$\bar{p}_v = p_v \exp(j\theta_{pv}) = m\bar{v}_v \quad (18)$$

where  $v = x, y, z$ . The momentum magnitudes can in general be written in four equivalent ways that can be used to pass to the four possible limiting symmetry states of spacetime namely, incoherent space and incoherent time, coherent space and incoherent time, incoherent space and coherent time, and coherent space and coherent time<sup>17,18,20</sup>

$$p_v = mv_v = m \sec \beta_{vv} \cos \beta_{tt}^v dv/dt_v \quad (19)$$

$$= m \csc \beta_{vv} \cos \beta_{tt}^v v d\theta_v/dt_v \quad (20)$$

$$= m \sec \beta_{vv} \sin \beta_{tt}^v t_v^{-1} dv/d\theta_t^v \quad (21)$$

$$= m \csc \beta_{vv} \sin \beta_{tt}^v v/t_v d\theta_v/d\theta_t^v \quad (22)$$

where  $v = x, y, z$ . The internal phase angles of the velocity and momentum components are written as<sup>17,18,20</sup>

$$\theta_{vv} = \theta_{pv} = \theta_v + \beta_{vv} - \theta_t^v - \beta_{tt}^v \quad (23)$$

The differentials of the complex number components of the single particle momentum can be written in two equivalent ways<sup>17,18,20</sup>

$$d\bar{p}_v = \sec \beta_{pvpv} dp_v \exp[j(\theta_{pv} + \beta_{pvpv})] \quad (24)$$

$$= \csc \beta_{pvpv} p_v d\theta_{pv} \exp[j(\theta_{pv} + \beta_{pvpv})] \quad (25)$$

where

$$\tan \beta_{pvpv} = p_v \partial \theta_{pv} / \partial p_v \quad (26)$$

$$\tan \beta_{vvvv} = v_v \partial \theta_{vv} / \partial v_v \quad (27)$$

and where

$$\beta_{pvpv} = \beta_{vvvv} \quad (28)$$

For high- $T_c$  materials the physically interesting spacetime states are the partial spacetime coherence case of nearly incoherent space and nearly incoherent time which described the normal state, and the case of coherent space and coherent time which describes the superconducting state. Ordinary metallic conductors correspond to incoherent space and incoherent time.

The angular speed of a body moving in the x,y plane in spacetime with broken internal symmetries is written as a complex number in internal space as<sup>17,18,20</sup>

$$\bar{\omega} = \omega \exp(j\theta_{\omega}) = d\bar{\phi}/d\bar{t}_{\phi} \quad (29)$$

where

$$\omega = \sec \beta_{\phi\phi} \cos \beta_{tt}^{\phi} d\phi/dt_{\phi} \quad (30)$$

$$= \csc \beta_{\phi\phi} \cos \beta_{tt}^{\phi} \phi d\theta_{\phi}/dt_{\phi} \quad (31)$$

$$= \sec \beta_{\phi\phi} \sin \beta_{tt}^{\phi} t_{\phi}^{-1} d\phi/d\theta_t^{\phi} \quad (32)$$

$$= \csc \beta_{\phi\phi} \sin \beta_{tt}^{\phi} \phi/t_{\phi} d\theta_{\phi}/d\theta_t^{\phi} \quad (33)$$

$$\theta_{\omega} = \theta_{\phi} + \beta_{\phi\phi} - \theta_t^{\phi} - \beta_{tt}^{\phi} \quad (34)$$

where

$$\tan \beta_{\phi\phi} = \phi \partial \theta_{\phi} / \partial \phi \quad (35)$$

$$\tan \beta_{tt}^{\phi} = t_{\phi} \partial \theta_t^{\phi} / \partial t_{\phi} \quad (36)$$

and where  $\bar{t}_{\phi}$  is the time associated with the azimuthal angle  $\phi$ .

The linear acceleration of a particle in broken symmetry spacetime is written as<sup>17,18,20</sup>

$$\bar{a}_v = a_v \exp(j\theta_{av}) = d\bar{v}_v/d\bar{t}_v = d^2\bar{v}/d\bar{t}_v^2 \quad (37)$$

where  $v = x, y, z, r$  or  $\rho$ . The case  $v = r$  corresponds to the linear radial acceleration in the x,y plane, so that

$$\bar{a}_r = a_r \exp(j\theta_{ar}) = d\bar{v}_r/d\bar{t}_r = d^2\bar{r}/d\bar{t}_r^2 \quad (38)$$

The magnitude and internal phase angle of the linear radial acceleration can be written in a number of different ways that are appropriate for the various states of coherence and incoherence of the space and time variations.<sup>20</sup> For the case of nearly incoherent space ( $\beta_{rr} \sim 0$ ) and nearly incoherent time ( $\beta_{tt}^r \sim 0$ ) the magnitude of the linear radial acceleration is written as<sup>20</sup>

$$a_r = \cos \beta_{tt}^r \sec \beta_{vrvr} d/dt(\cos \beta_{tt}^r \sec \beta_{rr} dr/dt) \quad (39)$$

$$\theta_{ar} = \theta_r + \beta_{rr} + \beta_{vrvr} - 2(\theta_t^r + \beta_{tt}^r) \quad (40)$$

where  $\beta_{tt}^r$  and  $\beta_{vrvr}$  are given by equations (9) and (27) as

$$\tan \beta_{tt}^r = t_r \partial \theta_t^r / \partial t_r \quad (41)$$

$$\tan \beta_{vrvr} = v_r \partial \theta_{vr} / \partial v_r \quad (42)$$

where  $\bar{t}_r$  is the time associated with the radial coordinate motion.

The magnitude of the linear radial acceleration term that is appropriate to describe nearly coherent space ( $\beta_{rr} \sim \pi/2$ ) and nearly coherent time ( $\beta_{tt}^r \sim \pi/2$ ) is given by the following expressions<sup>20</sup>

$$a_r = \sin \beta_{tt}^r (C_{rt}^2 + D_{rt}^2)^{1/2} / t_r \quad (43)$$

$$= \sin \beta_{tt}^r \csc \beta_{vrvr} C_{rt} / t_r$$

where

$$C_{rt} = \sin \beta_{tt}^r \csc \beta_{rr} r / t_r d\theta_r / d\theta_t^r [d\theta_r / d\theta_t^r - 1 + d/d\theta_t^r (\beta_{rr} - \beta_{tt}^r)] \quad (44)$$

$$D_{rt} = d/d\theta_t^r (\sin \beta_{tt}^r \csc \beta_{rr} r / t_r d\theta_r / d\theta_t^r) \quad (45)$$

$$\tan \beta_{vrvr} = C_{rt} / D_{rt} \quad (46)$$

$$\csc \beta_{vrvr} = (C_{rt}^2 + D_{rt}^2)^{1/2} / C_{rt} \quad (47)$$

and where the internal phase angle of the radial acceleration is

$$\theta_{ar} = \theta_r + \beta_{rr} + \beta_{vrvr} - 2(\theta_t^r + \beta_{tt}^r) \quad (48)$$

For an attractive force and a negative value of the linear radial acceleration it is convenient to write<sup>20</sup>

$$\bar{a}_r = a_r' \exp(j\theta_{ar}') \quad (49)$$

where

$$a_r' = -a_r \quad (49)$$

$$\theta_{ar}' = \theta_{ar} + \pi \quad (50)$$

$$= \theta_r + \beta_{rr} + \beta_{vrvr} - 2(\theta_t^r + \beta_{tt}^r) + \pi \quad (51)$$

For a positive acceleration



$$a_r^+ = a_r \quad (52)$$

$$\theta_{ar}^+ = \theta'_{ar} = \theta_r + \beta_{rr} + \beta_{vrvr}^+ - 2(\theta_t^r + \beta_{tt}^r) \quad (53)$$

$$\beta_{vrvr}^+ = \beta_{vrvr} + \pi \quad (54)$$

where  $\beta_{vrvr}$  is given by equation (47).

The general conditions for nearly incoherent spacetime are<sup>20</sup>

$$\beta_{tt}^v \sim 0 \quad \beta_{vv} \sim 0 \quad (55)$$

$$\theta_t^{vi} \sim \text{constant} \quad \theta_v^i \sim \text{constant} \quad (56)$$

where  $v$  and  $t_v$  are variables. The differentials of the spacetime coordinates for this special case are obtained from equations (7) and (8) as

$$dt_v^i = dt_v \exp(j\theta_t^{vi}) \quad (57)$$

$$dv^i = dv \exp(j\theta_v^i) \quad (58)$$

For this special case of nearly incoherent spacetime the momentum equations (19) and (23) give

$$p_v^i = mv_v^i = mdv/dt_v \quad (59)$$

$$\theta_{pv}^i = \theta_v^i - \theta_t^{vi} = \text{constant} \quad \beta_{pvpv}^i = 0 \quad (60)$$

and equation (24) becomes

$$dp_v^i = dp_v^i \exp[j(\theta_v^i - \theta_t^{vi})] \quad (61)$$

where  $v = x, y, z, r$  or  $\rho$  for the linear momenta. For rotational motion in nearly incoherent spacetime  $v = \phi$  and equations (55) and (56) give

$$\beta_{tt}^\phi \sim 0 \quad \beta_{\phi\phi} \sim 0 \quad (62)$$

$$\theta_t^{\phi i} \sim \text{constant} \quad \theta_\phi^i \sim \text{constant} \quad (63)$$

and equations (30) and (34) give

$$\omega^i = d\phi/dt_\phi \quad (64)$$

$$\theta_\omega^i = \theta_\phi^i - \theta_t^{\phi i} = \text{constant} \quad (65)$$

For nearly incoherent spacetime in the radial direction  $v = r$ ,  $\beta_{rr} \sim 0$ ,  $\beta_{tt}^r \sim 0$  and  $\beta_{vr} \sim 0$  so that equations (39) and (40) give

$$a_r^i = d^2 r / dt_r^2 \quad (66)$$

$$\theta_{ar}^i = \theta_r^i - 2\theta_t^{ri} = \text{constant} \quad (67)$$

For exactly incoherent spacetime all internal phase angles of the space and time coordinates have zero values, i.e.,  $\theta_v^i = 0$  and  $\theta_t^{vi} = 0$ .

For coherent spacetime<sup>20</sup>

$$\beta_{tt}^v = \pi/2 \quad \beta_{vv} = \pi/2 \quad (68)$$

$$t_v^c = t_v^c = \text{constant} \quad v = v^c = \text{constant} \quad (69)$$

where  $\theta_t^v$  and  $\theta_v$  are now variables. The differentials of the time and space coordinates are obtained from equations (7) and (8) as

$$d\bar{t}_v^c = j\bar{t}_v^c d\theta_t^v \quad (70)$$

$$d\bar{v}^c = j\bar{v}^c d\theta_v \quad (71)$$

For coherent spacetime equations (22) and (23) give

$$v_v^c = v^c / t^c d\theta_v / d\theta_t^v \quad (72)$$

$$\theta_{vv}^c = \theta_v - \theta_t^v \quad (73)$$

For the radial coordinate these equations become

$$\beta_{tt}^r = \pi/2 \quad \beta_{rr} = \pi/2 \quad (74)$$

$$t_r^c = t_r^c = \text{constant} \quad r = r^c = \text{constant} \quad (75)$$

$$d\bar{t}_r^c = j\bar{t}_r^c d\theta_t^r \quad d\bar{r}^c = j\bar{r}^c d\theta_r \quad (76)$$

$$v_r^c = r^c / t_r^c d\theta_r / d\theta_t^r \quad (77)$$

$$\theta_{vr}^c = \theta_r - \theta_t^r \quad (78)$$

The case of rotational motion in coherent spacetime is described by

$$\beta_{tt}^\phi = \pi/2 \quad \beta_{\phi\phi} = \pi/2 \quad (79)$$

$$t_{\phi}^c = t_{\phi}^c = \text{constant} \quad \phi = \phi^c = \text{constant} \quad (80)$$

$$d\bar{t}_{\phi}^c = j\bar{t}_{\phi}^c d\theta_{\phi}^c \quad d\bar{\phi}^c = j\bar{\phi}^c d\theta_{\phi}^c \quad (81)$$

and equations (33) and (34) give

$$\omega^c = \phi^c / t_{\phi}^c d\theta_{\phi}^c / d\theta_{\phi}^c \quad (82)$$

$$\theta_{\omega}^c = \theta_{\phi}^c - \theta_{\phi}^c \quad (83)$$

where  $\theta_{\phi}^c$  and  $\theta_{\phi}^c$  are now variables.

The complex number linear radial acceleration for a particle moving in coherent space and coherent time under the influence of an attractive force is obtained from equations (38) and (43) through (48) to be<sup>20</sup>

$$\bar{a}_r^c = a_r^c \exp(j\theta_{ar}^c) = \bar{r}^c / \bar{t}_r^{c2} (E_{rt}^c - jF_{rt}^c) \quad (84)$$

$$a_r^c = r^c / t_r^{c2} [(E_{rt}^c)^2 + (F_{rt}^c)^2]^{1/2} \quad (85)$$

$$\theta_{ar}^c = \theta_r^c + \beta_{vrvr}^c - 2\theta_t^r - \pi/2 \quad (86)$$

$$C_{rt}^c = r^c / t_r^c E_{rt}^c \quad D_{rt}^c = r^c / t_r^c F_{rt}^c \quad (87)$$

$$E_{rt}^c = d\theta_r / d\theta_t^r (d\theta_r / d\theta_t^r - 1) \quad E_{rt}^c \leq 0 \quad (88)$$

$$F_{rt}^c = d^2\theta_r / d\theta_t^{r2} \quad F_{rt}^c \geq 0 \quad (89)$$

$$\tan \beta_{vrvr}^c = E_{rt}^c / F_{rt}^c \quad (90)$$

$$\beta_{vrvr}^c = -\pi/2 + \delta_{rt} \quad (91)$$

$$\tan \delta_{rt} = F_{rt}^c / |E_{rt}^c| \quad (92)$$

Combining equations (86) and (91) gives the internal phase angle of the acceleration of a particle in coherent space and coherent time under the influence of an attractive force as

$$\theta_{ar}^c = \theta_r^c + \delta_{rt} - 2\theta_t^r - \pi \quad (93)$$

which has values in the neighborhood of  $-\pi$ . For an attractive force it is convenient to write the acceleration as<sup>20</sup>

$$\bar{a}_r^c = a_r^{c'} \exp(j\theta_{ar}^{c'}) \quad (94)$$

where

$$a_r^{c'} = -a_r^c = -r^c/t_r^{c2}[(E_{rt}^c)^2 + (F_{rt}^c)^2]^{1/2} \quad (95)$$

$$\theta_{ar}^{c'} = \theta_{ar}^c + \pi \quad (96)$$

$$= \theta_r + \beta_{vrvr}^c - 2\theta_t^r + \pi/2$$

$$= \theta_r + \delta_{rt} - 2\theta_t^r$$

where  $\theta_{ar}^{c'}$  is a small number.

For repulsive forces the acceleration of a particle in coherent spacetime is given by

$$\bar{a}_r^{c+} = \bar{r}^c/\bar{t}_r^{c2}(E_{rt}^{c+} - jF_{rt}^{c+}) \quad (97)$$

$$a_r^{c+} = r^c/t_r^{c2}[(E_{rt}^{c+})^2 + (F_{rt}^{c+})^2]^{1/2} \quad (98)$$

$$\theta_{ar}^{c+} = \theta_r + \beta_{vrvr}^{c+} - 2\theta_t^r - \pi/2 \quad (99)$$

with

$$E_{rt}^{c+} = -E_{rt}^c \quad E_{rt}^{c+} \geq 0 \quad (100)$$

$$F_{rt}^{c+} = -F_{rt}^c \quad F_{rt}^{c+} \leq 0 \quad (101)$$

$$\tan \beta_{vrvr}^{c+} = E_{rt}^{c+}/F_{rt}^{c+} \quad \tan \delta_{rt} = |F_{rt}^{c+}|/E_{rt}^{c+} \quad (102)$$

$$\beta_{vrvr}^{c+} = \beta_{vrvr}^c + \pi \quad (103)$$

$$= \pi/2 + \delta_{rt}$$

and

$$\theta_{ar}^{c+} = \theta_{ar}^{c'} = \theta_r + \delta_{rt} - 2\theta_t^r \quad (104)$$

so that  $\theta_{ar}^{c+}$  is a small angle. Finally if  $\bar{a}_r^c$  describes a negative acceleration

$$\bar{a}_r^{c+} = -\bar{a}_r^c \quad (105)$$

Note the different values of  $\beta_{vrvr}^c$  and  $\beta_{vrvr}^{c+}$  in equations (91) and (103) for attractive and repulsive forces.

For the special case of a solution of the form<sup>20</sup>

$$\theta_r = \alpha_r \theta_t^r + \beta_r \quad (106)$$

it follows that for coherent motion in the radial coordinate

$$F_{rt}^c = 0 \quad F_{rt}^{c+} = 0 \quad (107)$$

so that equations (91), (92), (102), (103) and (104) give for this case

$$\beta_{vr}^c = -\pi/2 \quad \beta_{vr}^{c+} = \pi/2 \quad \delta_{rt} = 0 \quad (108)$$

$$\theta_{ar}^{c'} = \theta_{ar}^{c+} = \theta_r - 2\theta_t^r \quad (109)$$

and therefore for this special case equations (24) and (25) give

$$d\bar{p}_r^c = j\bar{p}_r^c d\theta_{pr} \quad d\bar{v}_r^c = j\bar{v}_r^c d\theta_{vr} \quad (110)$$

$$d\bar{p}_r^{c+} = j\bar{p}_r^{c+} d\theta_{pr} \quad d\bar{v}_r^{c+} = j\bar{v}_r^{c+} d\theta_{vr} \quad (111)$$

where for coherent spacetime

$$\bar{p}_r^c = p_{rc} \exp(j\theta_{pr}) \quad (112)$$

with  $p_{rc} = \text{constant}$ . In the general case of

$$\theta_v = \alpha_v \theta_t^v + \beta_v \quad (113)$$

where  $v = x, y, z, r$  or  $\rho$ , it follows that for coherent spacetime and equations (24) and (25)

$$d\bar{p}_v^c = j\bar{p}_v^c d\theta_{pv} \quad d\bar{v}_v^c = j\bar{v}_v^c d\theta_{vv} \quad (114)$$

$$d\bar{p}_v^{c+} = j\bar{p}_v^{c+} d\theta_{pv} \quad d\bar{v}_v^{c+} = j\bar{v}_v^{c+} d\theta_{vv} \quad (115)$$

where

$$\bar{p}_v^c = p_{vc} \exp(j\theta_{pv}) \quad (116)$$

These expressions are used in Section 3 to evaluate the momentum space integrals that describe thermionic emission from the superconducting state of high- $T_c$  materials.

**2. NON-KEPLERIAN MOTION OF ELECTRONS IN COOPER PAIRS OF HIGH- $T_c$  SUPERCONDUCTORS.** This section describes the dynamical behavior of electrons in Cooper pairs which occur in the superconducting state of a high- $T_c$  material. A Cooper pair is a weakly bound system of two electrons which are held together by a weak attractive force which is mediated by the vibrations (phonons) of the crystalline

lattice of the high- $T_c$  superconductor.<sup>6-9</sup> The classical description of this bound system is given by the equation of motion of two electrons orbiting their center of mass<sup>21</sup>

$$\mu_e (a_r - r\omega^2) = F_i \quad (117)$$

where  $\mu_e = m_e/2$  = reduced mass of the electron,  $m_e$  = mass of the electron,  $r = 2a$  = relative distance between the two electrons,  $a$  = distance from center of mass to either electron, and where the linear radial acceleration term  $a_r$  is given by

$$a_r = d^2 r / dt^2 \quad (118)$$

and the angular frequency is given by

$$\omega = d\phi / dt \quad (119)$$

where  $\phi$  = azimuthal angle of an electron in it's orbit. Equation (117) simplifies for the case of a circular orbit with  $r = 2a$  = constant so that equation (117) becomes

$$\omega^2 = -F_i / (r\mu_e) = -F_i / (am_e) \quad (120)$$

where  $F_i < 0$  for an attractive force, and  $a$  = radius of orbit. Equation (120) shows that the orbital frequency varies inversely with the orbital radius even if the attractive interaction force  $F_i$  were independent of the relative distance between the electrons. In fact the pairing force between the two electrons of a Cooper pair is an attractive inverse square law given by<sup>1-9</sup>

$$F_i = -b/r^2 \quad (121)$$

where  $b > 0$ , so that equation (120) becomes

$$\omega^2 = b/(\mu_e r^3) = b/(4m_e a^3) \quad (122)$$

which is essentially Kepler's law of central field motion.<sup>21</sup> This section considers only attractive forces.

In the normal and superconducting states of a high- $T_c$  material the space and time coordinates of the electrons in Cooper pairs exhibit broken internal symmetries and must be written in the form of equations (4) and (5). Therefore for a high- $T_c$  material the equation of motion of the electrons in a Cooper pair are written as

$$\mu_e (\bar{a}_r - \bar{r}\bar{\omega}^2) = \bar{F}_i \quad (123)$$

where  $\bar{a}_r$  = complex number linear radial acceleration given by equations (38) through (48),  $\bar{\omega}$  = complex number angular speed given by equations (29) through (34), and where  $\bar{r}$  = complex number relative distance between the two electrons

and  $\bar{\phi}$  = complex number azimuthal angle of the electrons which are written as

$$\bar{r} = r \exp(j\theta_r) \quad \bar{\phi} = \phi \exp(j\theta_\phi) \quad (124)$$

The associated complex number time coordinates are given by

$$\bar{t}_r = t_r \exp(j\theta_t^r) \quad \bar{t}_\phi = t_\phi \exp(j\theta_t^\phi) \quad (125)$$

The complex number electron-electron interaction force is written as

$$\bar{F}_i = F_i \exp(j\theta_{Fi}) \quad (126)$$

For an attractive inverse square force

$$\begin{aligned} \bar{F}_i &= -\bar{b}/\bar{r}^2 \\ F_i &= -b/r^2 \quad \theta_{Fi} = -2\theta_r + \theta_b \end{aligned} \quad (128)$$

where  $\bar{b}$  is taken to be

$$\bar{b} = \bar{g}e^2/(4\pi) \quad (129)$$

where  $\bar{g}$  = complex number electron-phonon interaction constant, and  $e$  = electron charge. These complex number parameters are written as

$$\bar{b} = b \exp(j\theta_b) \quad \bar{g} = g \exp(j\theta_g) \quad (130)$$

so that equation (129) can be written as

$$b = ge^2/(4\pi) \quad \theta_b = \theta_g \quad (131)$$

where  $b > 0$  and  $g > 0$ .

The total angular momentum of the two electrons in a Cooper pair is a complex number in internal space for spacetime with broken internal symmetries and is written as

$$\bar{L}_z = L_z \exp(j\theta_{Lz}) = \mu_e \bar{r}^2 \bar{\omega} \quad (132)$$

where  $\bar{r}$  = complex number relative radial distance between the two electrons, and  $\bar{\omega}$  = complex number angular speed which is given by equations (29) through (36). The law of the conservation of angular momentum for this case is written as<sup>21</sup>

$$\bar{L}_z = \mu_e \bar{r}^2 \bar{\omega} = \text{constant} \quad (133)$$

$$L_z = \mu_e r^2 \omega = \text{constant} \quad (134)$$

$$\theta_{Lz} = 2\theta_r + \theta_\omega = \text{constant} \quad (135)$$

where  $\omega$  is given by equations (30) through (33) and  $\theta_\omega$  is given by equation (34). Combining conditions (34) and (135) gives

$$\theta_{Lz} = 2\theta_r + \theta_\phi + \beta_{\phi\phi} - \theta_t^\phi - \beta_{tt}^\phi = \text{constant} \quad (136)$$

where  $\beta_{tt}^\phi$  is given by equation (36).

A full solution to equation (123) requires that the real and imaginary parts of this equation be obtained and solved jointly. This procedure leads to complicated equations which are not easily solved. A simpler, but approximate, procedure is to assume that the internal phase angles of each term in equation (123) are equal. This gives the following two equations for the case where the linear radial acceleration term is positive

$$\mu_e(a_r^+ - \omega^2 r) = F_i \quad (137)$$

$$\theta_{ar}^+ = \theta_r + 2\theta_\omega = \theta_{Fi} \quad (138)$$

where

$$a_r^+ = a_r \quad \theta_{ar}^+ = \theta_{ar} + \pi \quad (139)$$

and the following two equations for the case when the linear radial acceleration term is negative

$$\mu_e(a_r' - \omega^2 r) = F_i \quad (140)$$

$$\theta_{ar}' = \theta_r + 2\theta_\omega = \theta_{Fi} \quad (141)$$

where for this case

$$a_r' = -a_r \quad \theta_{ar}' = \theta_{ar} + \pi \quad (142)$$

In both cases  $\theta_{ar}^+ = \theta_{ar}' = \text{small numbers}$ , and  $a_r$  is given by equation (43) while  $\omega$  is given by equation (30) for the conditions of interest in this paper. Combining equations (53) and (138) gives for a positive linear acceleration term

$$\begin{aligned} \theta_{Fi} &= \theta_r + \beta_{rr} + \beta_{vrvr}^+ - 2(\theta_t^r + \beta_{tt}^r) \\ &= \theta_r + 2\theta_\omega \end{aligned} \quad (143)$$

where

$$\beta_{vrvr}^+ = \beta_{vrvr} + \pi \quad (144)$$

For a negative linear acceleration term equations (51) and (141) give

$$\begin{aligned} \theta_{Fi} &= \theta_r + \beta_{rr} + \beta_{vrvr} - 2(\theta_t^r + \beta_{tt}^r) + \pi \\ &= \theta_r + 2\theta_\omega \end{aligned} \quad (145)$$



Equations (143) through (145) are valid for the case of nearly coherent space and time in the radial direction. Combining equations (135) and (138) or (141) gives

$$\theta_{Fi} = 2\theta_{Lz} - 3\theta_r \quad (146)$$

$$= 1/2(\theta_{Lz} + 3\theta_\omega) \quad (147)$$

$$= 1/2[\theta_{Lz} + 3(\theta_\phi + \beta_{\phi\phi} - \theta_t^\phi - \beta_{tt}^\phi)] \quad (148)$$

The quantity  $\theta_{ar}^+ = \theta_{ar}' = \theta_{Fi}$  enters into the calculation of the normalized superconductivity energy gap for high- $T_c$  superconductors.<sup>14</sup>

For an attractive inverse square law described by equation (127) it follows from equations (51), (128) and (141) or equivalently from equations (53), (54), (128) and (138) that

$$3\theta_r + \beta_{rr} + \beta_{vrvr} - 2(\theta_t^r + \beta_{tt}^r) + \pi = \theta_b \quad (149)$$

$$3\theta_r + 2\theta_\omega = \theta_b \quad (150)$$

Equation (149) can also be written as

$$3\theta_r + \beta_{rr} + \beta_{vrvr}^+ - 2(\theta_t^r + \beta_{tt}^r) = \theta_b \quad (151)$$

Equations (135), (146) and (150) yield

$$\theta_r = 2\theta_{Lz} - \theta_b \quad \theta_{Fi} = 3\theta_b - 4\theta_{Lz} \quad (152)$$

$$\theta_\omega = 2\theta_b - 3\theta_{Lz} \quad (153)$$

where  $\theta_\omega$  is given by equation (34). Therefore within the approximations given in equations (138) and (141) it follows that  $\theta_r$ ,  $\theta_\omega$  and  $\theta_{Fi}$  are constants. Equations (34), (151) and (153) show that

$$\begin{aligned} 2(\theta_\phi + \beta_{\phi\phi} - \theta_t^\phi - \beta_{tt}^\phi) &= \beta_{rr} + \beta_{vrvr} - 2(\theta_t^r + \beta_{tt}^r) + \pi \quad (153A) \\ &= \beta_{rr} + \beta_{vrvr}^+ - 2(\theta_t^r + \beta_{tt}^r) \\ &= 2(2\theta_b - 3\theta_{Lz}) \\ &= \text{constant} \end{aligned}$$

which is an approximate equation relating the radial and azimuthal internal phase angles of space and time.

The case where the radial acceleration is totally coherent in space and time is described by the conditions

$$\beta_{rr} = \pi/2 \quad \beta_{tt}^r = \pi/2 \quad (154)$$

For this case equations (143) and (145) become

$$\begin{aligned} \theta_{Fi} &= \theta_r + \beta_{vrvr}^c - 2\theta_t^r + \pi/2 \\ &= \theta_r + 2\theta_\omega \end{aligned} \quad (155)$$

where  $\beta_{vrvr}^c$  is given by equation (91) so that

$$\begin{aligned} \theta_{Fi} &= \theta_r - 2\theta_t^r + \delta_{rt} \\ &= \theta_r + 2\theta_\omega \end{aligned} \quad (156)$$

or

$$\begin{aligned} 2\theta_\omega &= \delta_{rt} - 2\theta_t^r \\ &= 2(2\theta_b - 3\theta_{Lz}) = \text{constant} \end{aligned} \quad (157)$$

where  $\theta_\omega = \text{constant}$  given by equation (153). Equation (156) is seen to agree with equations (96) and (104). Combining equations (146) and (156) gives

$$\theta_{Lz} = 2\theta_r + \delta_{rt}/2 - \theta_t^r \quad (158)$$

which is valid for the special case of coherent space and time in the radial direction. For a central force equations (128) and (156), or equivalently equation (149) gives

$$\theta_b = 3\theta_r + \delta_{rt} - 2\theta_t^r \quad (159)$$

where in these equations  $\theta_r = \text{constant}$  given by equation (152). Equations (158) and (159) can be used to determine  $\theta_t^r$  as

$$\begin{aligned} \theta_t^r &= 3\theta_{Lz} - 2\theta_b + \delta_{rt}/2 \\ &= 2\theta_r - \theta_{Lz} + \delta_{rt}/2 \\ &= 1/2(3\theta_r - \theta_b + \delta_{rt}) \end{aligned} \quad (160)$$

where  $\delta_{rt}$  is given by (88), (89) and (92), so that  $\theta_t^r$  is also a constant.

The following conditions are valid when the space and time coordinates are coherent in the radial direction but the angular space and time coordinates are nearly incoherent

$$\beta_{rr} = \pi/2 \quad \beta_{tt}^r = \pi/2 \quad (161)$$

$$\beta_{\phi\phi} \sim 0 \quad \beta_{tt}^\phi \sim 0 \quad (162)$$

$$\theta_\phi \sim \theta_\phi^i \quad \theta_t^\phi \sim \theta_t^{\phi i} \quad (163)$$

where  $\theta_\phi^i$  and  $\theta_t^{\phi i}$  are constants which would have zero values for the case of exact incoherence of the azimuthal space and time coordinates. Combining equations (30), (34) and (162) gives

$$\omega = d\phi/dt \quad (164)$$

$$\theta_\omega = \theta_\phi^i - \theta_t^{\phi i} \quad (165)$$

Combining equations (7), (8), (11), (12), (161) and (162) gives

$$d\bar{r} = j\bar{r}d\theta_r \quad (166)$$

$$d\bar{t}_r = j\bar{t}_r d\theta_t^r \quad (167)$$

$$d\bar{\phi} = d\phi \exp(j\theta_\phi^i) \quad (168)$$

$$d\bar{t}_\phi = dt_\phi \exp(j\theta_t^{\phi i}) \quad (169)$$

Then combining equations (157) and (165) gives

$$\delta_{rt} - 2\theta_t^r = 2(\theta_\phi^i - \theta_t^{\phi i}) \quad (170)$$

Introducing equation (160) into equation (170) gives

$$\delta_{rt} - 2(3\theta_{Lz} - 2\theta_b + \delta_{rt}/2) = 2(\theta_\phi^i - \theta_t^{\phi i}) \quad (171)$$

which can be rewritten as

$$2\theta_b - 3\theta_{Lz} = \theta_\phi^i - \theta_t^{\phi i} \quad (172)$$

Equations (152), (156), (160) and (172) give

$$\begin{aligned} \theta_{Fi} &= \theta_r - 2\theta_t^r + \delta_{rt} \\ &= \theta_r + 2(2\theta_b - 3\theta_{Lz}) \\ &= \theta_r + 2(\theta_\phi^i - \theta_t^{\phi i}) \\ &= 3\theta_b - 4\theta_{Lz} \end{aligned} \quad (173)$$

where  $\theta_r$  is given by equation (152). If the constants  $\bar{b}$  and  $\bar{L}_z$  are related in the following manner

$$\bar{b}^2 = f\bar{L}_z^3 \quad (174)$$

or equivalently as

$$b^2 = fL_z^3 \quad 2\theta_b = 3\theta_{Lz} \quad (175)$$

where  $f$  is a real number constant, then it follows from equations (153), (165), (172) and (175) that

$$\theta_\phi^i = \theta_t^{\phi i} \quad \theta_\omega = 0 \quad (176)$$

while equations (160) and (173) give for this special case

$$\delta_{rt} = 2\theta_t^r \quad (177)$$

$$\theta_{Fi} = \theta_r = \theta_b/3 = \theta_{Lz}/2 \quad (178)$$

which is valid for the general case  $\delta_{rt} \neq 0$ .

Consider now the possibility of the case where the space and time coordinates are coherent in both the radial and the azimuthal directions which is described by

$$\beta_{rr} = \pi/2 \quad \beta_{tt}^r = \pi/2 \quad \beta_{\phi\phi} = \pi/2 \quad \beta_{tt}^\phi = \pi/2 \quad (179)$$

where now  $\theta_r$ ,  $\theta_t^r$ ,  $\theta_\phi$  and  $\theta_t^\phi$  are variables. For this case equations (143) and (144) become

$$\beta_{rvrv}^c - 2\theta_t^r + \pi/2 = 2(\theta_\phi - \theta_t^\phi) \quad (180)$$

and then using equation (91) gives

$$\delta_{rt} - 2\theta_t^r = 2(\theta_\phi - \theta_t^\phi) \quad (181)$$

which is similar in form to equation (170) except that the right hand side of equation (170) is a constant while the right hand side of equation (181) is a variable because for this case equations (33) and (34) give

$$\omega^c = \phi^c/t_\phi^c \, d\theta_\phi/d\theta_t^\phi \quad (182)$$

$$\theta_\omega^c = \theta_\phi - \theta_t^\phi \quad (183)$$

which are variables. But within the approximation used in equations (157) and (160) it follows that for incoherent spacetime in the azimuthal direction

$$\delta_{rt} - 2\theta_t^r = 2(2\theta_b - 3\theta_{Lz}) \quad (184)$$

which is a constant.

For the case where space and time are coherent in the radial direction it is often expedient to assume a linear relationship between  $\theta_r$  and  $\theta_t^r$  of the form<sup>20</sup>

$$\theta_r = \alpha_r \theta_t^r + \beta_r \quad \alpha_r \leq 1 \quad (185)$$

where the coefficient  $\alpha_r$  is determined by the nature of the pairing force which is related to the atomic structure of the crystal lattice of the high- $T_c$  material. Combining equations (88), (89) and (185) gives

$$E_{rt}^c = \alpha_r (\alpha_r - 1) \quad F_{rt}^c = 0 \quad (186)$$

while equations (90) and (91) give

$$\beta_{vrvr}^c = -\pi/2 \quad \delta_{rt} = 0 \quad (187)$$

and equations (96) and (104) become

$$\theta_{ar}^{c+} = \theta_{ar}^{c'} = \theta_r - 2\theta_t^r \quad (188)$$

Combining equations (157) and (187) gives for this special case of coherence of space and time in the radial direction

$$\theta_\omega = -\theta_t^r \quad (189)$$

while equation (156) becomes

$$\begin{aligned} \theta_{Fi} &= \theta_r - 2\theta_t^r \\ &= \theta_r + 2\theta_\omega \\ &= 2\theta_{Lz} - 3\theta_r \end{aligned} \quad (190)$$

where equation (158) gives

$$\theta_{Lz} = 2\theta_r - \theta_t^r \quad (191)$$

For an inverse square electron-electron pairing force equations (159) and (187) give

$$\theta_b = 3\theta_r - 2\theta_t^r \quad (192)$$

Equations (185) and (192) give for an inverse square pairing force

$$\alpha_r = d\theta_r / d\theta_t^r = 2/3 \quad (193)$$

so that equation (186) gives for this case

$$E_{rt}^c = \alpha_r(\alpha_r - 1) = -2/9 \quad F_{rt}^c = 0 \quad (194)$$

Combining equations (152), (153), (160), (186) and (187) gives the following results for an inverse square type of pairing force and the assumption of the validity of equation (185) which has  $\delta_{rt} = 0$

$$\theta_t^r = 3\theta_{Lz} - 2\theta_b \quad \theta_{Fi} = 3\theta_b - 4\theta_{Lz} \quad (195)$$

$$\theta_r = 2\theta_{Lz} - \theta_b \quad (196)$$

$$\theta_\omega = 2\theta_b - 3\theta_{Lz} \quad (197)$$

Within the approximations made in this section all internal phase angles are constants. A further example of this is obtained from equations (188) and (190) which give

$$\theta_{ar}^{c'} = 3\theta_b - 4\theta_{Lz} \quad (198)$$

$$\theta_{Fi} = 3\theta_b - 4\theta_{Lz} \quad (199)$$

In the special case when equation (175) is valid it follows from equations (195) through (199) that

$$\theta_r = \theta_{Lz}/2 = \theta_b/3 \quad (200)$$

$$\theta_t^r = 0 \quad \theta_\omega = 0 \quad \delta_{rt} = 0 \quad (201)$$

$$\theta_{ar}^{c'} = \theta_{Fi} = \theta_b/3 = \theta_{Lz}/2 = \theta_r \quad (202)$$

The intrinsic signs of  $\theta_{Lz}$  and  $\theta_b$  are negative because the intrinsic sign of  $\theta_r$  is negative.<sup>17,18</sup>

For the case of coherent spacetime in the radial direction of the orbiting electrons of a Cooper pair, equation (85) gives the magnitude of the relative radial acceleration of the two electrons as

$$a_r^c = r^c / t_r^{c2} [(E_{rt}^c)^2 + (F_{rt}^c)^2]^{1/2} \quad (203)$$

where  $r^c$  = constant magnitude of the relative separation of the two electrons,  $t_r^c$  = constant magnitude of the time for the radial direction and where  $E_{rt}^c$  and  $F_{rt}^c$  are given by equations (88) and (89). The acceleration of an electron relative to the center of mass of the electron pair is given by  $a_r^c/2$ . The value of  $t_r^c$  can be taken to be a characteristic time of the electron pair system - the Bohr time  $t_B$  of the electron pair orbiting a lattice charge of  $Z = 2$

$$t_r^c = t_B = \hbar^3 / (m_e e^4) = \hbar / (2|E_B|) \quad (204)$$

where  $\hbar = h/(2\pi)$ ,  $h$  = Planck's constant,  $E_B = -e^2/(2a_0)$  = energy of an electron in the ground state Bohr orbit,  $a_0$  = Bohr radius, and  $m_e$  = mass of the electron. The factor of 2 in equation (204) results from the product of a factor of 4 which results from  $Z^2 = 4$  and a factor of 1/2 which results from the reduced mass of the electron pair  $\mu_e = m_e/2$ . Then the equation of motion for the electron pair with coherent spacetime coordinates in the radial direction and incoherent spacetime coordinates in the azimuthal direction is obtained from equation (137) to be

$$[(E_{rt}^c)^2 + (F_{rt}^c)^2]^{1/2}/t_B^2 - \omega^2 = F_i/(\mu_e r^c) \quad (205)$$

where  $\mu_e = m_e/2$  = reduced mass of the electron, and where  $\omega = d\phi/dt$  = incoherent angular speed of the electrons. For the case of a weak electron-electron interaction force  $F_i \sim 0$ , and equation (205) becomes

$$\omega = t_B^{-1} [(E_{rt}^c)^2 + (F_{rt}^c)^2]^{1/4} \quad (206)$$

The incoherent angular speed of the electron pair required for equilibrium is therefore independent of the relative distance  $r^c$  between the two electrons, and represents non-Keplerian motion. For comparison equation (122) represents Keplerian motion. Equation (206) is only valid for a weak pairing force.

The equilibrium radius  $a_c$  of the orbit of the electrons in a Cooper pair is obtained from equations (134), (204) and (206) as

$$\begin{aligned} a_c^2 &= \ell_z / (m_e \omega) \\ &= \ell_z t_B m_e^{-1} [(E_{rt}^c)^2 + (F_{rt}^c)^2]^{-1/4} \end{aligned} \quad (207)$$

where  $\ell_z = L_z/2$  = angular momentum of an electron, where  $L_z$  = total angular momentum of the two electrons given by equation (134), and where  $a_c = r^c/2$  where  $r^c$  = equilibrium distance between the two electrons in radial coherent spacetime. Equations (206) and (207) show that both  $\omega$  and  $a_c$  are independent of the nature of the weak pairing force  $F_i$ . The angular momentum of an electron is quantized in the usual manner<sup>22</sup>

$$\ell_z = m\hbar \quad (208)$$

where  $m$  = magnetic quantum number which can take the values  $m = 0, \pm 1, \pm 2, \pm 3, \dots$ . Combining equations (207) and (208) gives the quantized circular orbits of the electron in a Cooper pair of a high- $T_c$  superconductor as

$$a_{cm}^2 = m\hbar t_B m_e^{-1} [(E_{rt}^c)^2 + (F_{rt}^c)^2]^{-1/4} \quad (209)$$

where  $a_{cm}$  = orbit radius corresponding to the magnetic quantum number  $m$ .

For high- $T_c$  superconductors  $E_{rt}^c$  and  $F_{rt}^c$  can be obtained from equation (186) so that equations (206) and (209) become respectively

$$\omega = t_B^{-1} [\alpha_r (1 - \alpha_r)]^{1/2} \quad (210)$$

$$a_{cm}^2 = m \hbar t_B m_e^{-1} [\alpha_r (1 - \alpha_r)]^{-1/2} \quad (211)$$

Combining equations (193), (210) and (211) gives

$$\omega = \sqrt{2}/3 t_B^{-1} = \sqrt{2}/3 m_e^4 / \hbar^3 \quad (212)$$

$$a_{cm}^2 = 3/\sqrt{2} m \hbar t_B m_e^{-1} = 3/\sqrt{2} m \hbar^4 / (m_e e^2)^2 \quad (213)$$

Equation (213) gives the radius of the  $m$ 'th orbit as

$$\begin{aligned} a_{cm} &= (3/\sqrt{2})^{1/2} m^{1/2} a_0 \\ &= 1.456 m^{1/2} a_0 \end{aligned} \quad (214)$$

where the Bohr radius  $a_0$  is given by<sup>22</sup>

$$a_0 = \hbar^2 / (m_e e^2) \quad (215)$$

The energy of a bound electron in the pair can be obtained from equation (214) as

$$\begin{aligned} E_m &= - e^2 / (2 a_{cm}) \\ &= - 0.343 m^{-1/2} e^2 / a_0 = 0.687 m^{-1/2} E_B \end{aligned} \quad (216)$$

where  $m$  = magnetic quantum number and  $E_B = - e^2 / (2 a_0)$  is the energy of an electron in a Bohr orbit. For comparison, the standard expressions for the radius and energy of a Bohr atom with  $Z = 1$  is given by

$$a_n = n^2 a_0 \quad (217)$$

$$\begin{aligned} E_n &= - e^2 / (2 a_n) = - e^2 / (2 n^2 a_0) \\ &= n^{-2} E_B \end{aligned} \quad (218)$$

where  $n = 1, 2, 3, \dots$  is the principal quantum number.

**3. THERMIONIC EMISSION FROM HIGH- $T_c$  MATERIALS.** Thermionic emission refers to electrons emitted from the surface of solids that are heated to some specified temperature. This phenomenon is essentially a tunneling process because there is an energy barrier of a few eV that prevents most of the electron gas in a solid from escaping.<sup>23-25</sup> At a finite temperature however some electrons in the solid have sufficient energy to penetrate the barrier and escape from the solid surface. The thermionic emission from high- $T_c$  superconductors



is relatively small but it contains significant information about the electrons that occur in weakly bound Cooper pairs. Therefore it is important to develop a theory of the thermionic emission from the normal and superconducting states of high- $T_c$  materials and to compare the predicted thermionic emission current with measured values. In this way the values of the internal phase angles of the electron momenta and the internal phase angles of the space and time coordinates of the electrons can be determined, and therefore the degree of coherence of spacetime can be determined for high- $T_c$  materials.

#### A. Richardson-Dushman Thermionic Emission Equation.

Before considering the thermionic emission from high- $T_c$  materials a brief review is given of the standard theory of thermionic emission from ordinary metals.<sup>23-25</sup> The conventional picture of electrons in a metal is that they form a Fermi gas whose distribution function is given by<sup>23-25</sup>

$$f = \{\exp[(\epsilon - \mu)/(kT)] + 1\}^{-1} \quad (219)$$

where  $\epsilon$  = kinetic energy of an electron,  $\mu$  = chemical potential,  $k$  = Boltzmann constant and  $T$  = absolute temperature. The electron kinetic energy is written as

$$\epsilon = (p_x^2 + p_y^2 + p_z^2)/(2m_e) \quad (220)$$

where  $p_x$ ,  $p_y$  and  $p_z$  = components of electron momentum, and where  $m_e$  = electron mass. The  $x$  direction is taken as the coordinate axis that is normal to the surface of the metal. For the electrons involved in thermionic emission, the kinetic energy per electron must be greater than a critical value given by

$$\epsilon_k = p_{xk}^2/(2m_e) = 1/2 m_e v_{xk}^2 = \mu + e\phi \quad (221)$$

where  $\epsilon_k$ ,  $p_{xk}$  and  $v_{xk}$  = critical value of the electron kinetic energy, critical electron momentum normal to the surface, and critical electron velocity normal to the surface of a metal,  $e$  = electron charge, and  $\phi$  = work function such that  $e\phi$  = energy required to remove an electron which is at the top of the Fermi sea of electrons. Therefore the critical value of the electron momentum normal to the metal surface that is needed to just eject an electron from the metal surface with zero velocity is given by

$$p_{xk} = [2m_e(\mu + e\phi)]^{1/2} \quad (222)$$

For metals the work function is of the order of  $\phi > 1$  volt, so that in general for thermionic emission equations (220) through (222) give

$$\epsilon > \epsilon_k \gg \mu \quad (223)$$

and therefore as far as the thermionic electrons are concerned the distribution function given in equation (219) can be written as

$$\begin{aligned} f &\sim \exp[-(\epsilon - \mu)/(kT)] \\ &= \exp[-(p_x^2 + p_y^2 + p_z^2 - 2m_e\mu)/(2m_e kT)] \end{aligned} \quad (224)$$

The electron number density of a Fermi gas is written as<sup>23-25</sup>

$$dn_e = f dn_p = 2/h^3 f dp_x dp_y dp_z \quad (225)$$

where  $n_e$  = electron number density and  $n_p$  = number density of momentum states.

The thermionic emission current can be written as<sup>23-25</sup>

$$\begin{aligned} I_x &= e \int v_x dn_e = e/m_e \int p_x dn_e \\ &= 2e/(m_e h^3) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{p_{xk}}^{\infty} p_x f dp_x dp_y dp_z \\ &= 2e/(m_e h^3) J_x J_y J_z \end{aligned} \quad (226)$$

where

$$J_x = \int_{p_{xk}}^{\infty} \exp[-(p_x^2 - 2m_e \mu)/(2m_e kT)] p_x dp_x \quad (227)$$

$$J_y = \int_{-\infty}^{\infty} \exp[-p_y^2/(2m_e kT)] dp_y \quad (228)$$

$$J_z = \int_{-\infty}^{\infty} \exp[-p_z^2/(2m_e kT)] dp_z \quad (229)$$

The integrals can be evaluated from tables with the result that<sup>26</sup>

$$J_x = m_e kT \exp[-e\phi/(kT)] \quad (230)$$

$$J_y = J_z = (2\pi m_e kT)^{1/2} \quad (231)$$

where the term  $-e\phi/(kT)$  in equation (230) comes from the lower limit of integration  $p_{xk}$  in equation (222) and (227). Combining equations (226), (230) and (231) gives the well known Richardson-Dushman thermionic emission equation<sup>23-25</sup>

$$I_x = A_o T^2 \exp[-e\phi/(kT)] \quad (232)$$

where

$$A_o = 4\pi m_e e k^2/h^3 \quad (233)$$

is the Richardson-Dushman constant.

#### B. Thermionic Emission from High- $T_c$ Materials.

Because the factor  $6/\pi$ , which arises from the assumption of coherent time,

is responsible for the relatively high values of the normalized superconducting energy gap given by equation (3), it is reasonable to assume that the electrons in the Cooper pairs are weakly bound.<sup>13,14</sup> Therefore as a first approximation to calculating the thermionic emission current for high- $T_c$  materials it will be assumed that the electrons form a noninteracting Fermi gas and that the basic ideas of the Richardson-Dushman calculation can be utilized. For the normal state of a high- $T_c$  material the electrons are assumed to move in a partially coherent spacetime. Then the complex number generalization of equations (219) through (225) are<sup>14</sup>

$$\bar{f} = \{\exp[(\bar{\epsilon} - \bar{\mu})/(kT)] + 1\}^{-1} \quad (234)$$

$$\begin{aligned} &\sim \exp[-(\bar{\epsilon} - \bar{\mu})/(kT)] \\ &= \exp[-(\bar{p}_x^2 + \bar{p}_y^2 + \bar{p}_z^2 - 2m_e \bar{\mu})/(2m_e kT)] \end{aligned}$$

$$\bar{\epsilon} = (\bar{p}_x^2 + \bar{p}_y^2 + \bar{p}_z^2)/(2m_e) \quad (235)$$

$$\bar{\epsilon}_k = \bar{p}_{xk}^2/(2m_e) = 1/2m_e \bar{v}_{xk}^2 = \bar{\mu} + e\bar{\phi} \quad (236)$$

$$\bar{p}_{xk} = [2m_e(\bar{\mu} + e\bar{\phi})]^{1/2} \quad (237)$$

$$d\bar{n}_e = \bar{f} d\bar{n}_p = 2/h^3 \bar{f} d\bar{p}_x d\bar{p}_y d\bar{p}_z \quad (238)$$

where  $\bar{p}_x$ ,  $\bar{p}_y$  and  $\bar{p}_z$  are represented as in equation (18) and where  $\bar{p}_{xk}$ ,  $\bar{\mu}$  and  $\bar{\phi}$  are written as

$$\bar{p}_{xk} = p_{xk} \exp(j\theta_{pxk}) \quad (239)$$

$$\bar{\mu} = \mu \exp(j\theta_\mu) \quad (240)$$

$$\bar{\phi} = \phi \exp(j\theta_\phi) \quad (241)$$

The values of  $p_{xk}$  and  $\theta_{pxk}$  are obtained from equation (237) by writing

$$p_{xk}^2 \exp(j2\theta_{pxk}) = 2m_e [\mu \exp(j\theta_\mu) + e\phi \exp(j\theta_\phi)] \quad (242)$$

whose real and imaginary parts are

$$p_{xk}^2 \cos(2\theta_{pxk}) = 2m_e (\mu \cos \theta_\mu + e\phi \cos \theta_\phi) \quad (243)$$

$$p_{xk}^2 \sin(2\theta_{pxk}) = 2m_e (\mu \sin \theta_\mu + e\phi \sin \theta_\phi) \quad (244)$$

Equations (243) and (244) give

$$\tan(2\theta_{pxk}) = (\mu \sin \theta_\mu + e\phi \sin \theta_\phi)(\mu \cos \theta_\mu + e\phi \cos \theta_\phi)^{-1} \quad (245)$$

$$p_{xk}^2 = 2m_e [\mu^2 + e^2 \phi^2 + 2\mu e\phi \cos(\theta_\mu - \theta_\phi)]^{1/2} \quad (246)$$

The momentum constant  $\bar{p}_{xk}$  appears as a lower integration limit in the complex integral that describes thermionic emission.

The complex number thermionic emission current is given by

$$\begin{aligned}\bar{I}_x &= e \int \bar{v}_x d\bar{n}_e = e/m_e \int \bar{p}_x d\bar{n}_e \\ &= 2e/(m_e h^3) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{\bar{p}_{xk}}^{\infty} \bar{p}_x \bar{f} d\bar{p}_x d\bar{p}_y d\bar{p}_z \\ &= 2e/(m_e h^3) \bar{J}_x \bar{J}_y \bar{J}_z\end{aligned}\quad (247)$$

where

$$\bar{J}_x = \int_{\bar{p}_{xk}}^{\infty} \exp[-(\bar{p}_x^2 - 2m_e \bar{u})/(2m_e kT)] \bar{p}_x d\bar{p}_x \quad (248)$$

$$\bar{J}_y = \int_{-\infty}^{\infty} \exp[-\bar{p}_y^2/(2m_e kT)] d\bar{p}_y \quad (249)$$

$$\bar{J}_z = \int_{-\infty}^{\infty} \exp[-\bar{p}_z^2/(2m_e kT)] d\bar{p}_z \quad (250)$$

The complex number thermionic current in equation (247) can be written as

$$\bar{I}_x = I_x \exp(j\theta_{Ix}) \quad (251)$$

and the integrals in equations (248) through (250) can be written as

$$\bar{J}_x = J_x \exp(j\theta_{Jx}) \quad (252)$$

$$\bar{J}_y = J_y \exp(j\theta_{Jy}) \quad (253)$$

$$\bar{J}_z = J_z \exp(j\theta_{Jz}) \quad (254)$$

Comparing equation (247) and (251) through (254) gives the thermionic emission current as

$$I_x = 2e/(m_e h^3) J_x J_y J_z \quad (255)$$

$$\theta_{Ix} = \theta_{Jx} + \theta_{Jy} + \theta_{Jz} \quad (256)$$

The measured thermionic emission current is given by the real part of equation (251) as

$$\begin{aligned}I_{xm} &= I_x \cos \theta_{Ix} \\ &= 2e/(m_e h^3) J_x J_y J_z \cos(\theta_{Jx} + \theta_{Jy} + \theta_{Jz})\end{aligned}\quad (257)$$

Therefore the problem at hand is to calculate  $J_x, J_y, J_z$  and  $\theta_{Jx}, \theta_{Jy}, \theta_{Jz}$ . In order to do this the real and imaginary parts of the integrals in equations (248) through (250) must be calculated.

The integrals in equations (248) through (250) can be rewritten using equations (24) and (25) as

$$\bar{J}_x = \int_{\bar{p}_{xk}}^{\infty} \bar{p}_x \bar{J}_{px} d\bar{p}_x \quad (258)$$

$$= \int_{\bar{p}_{xk}}^{\infty} p_x J_{px} \sec \beta_{pxpx} \exp[j(\theta_{Jpx} + 2\theta_{px} + \beta_{pxpx})] dp_x \quad (259)$$

$$= \int_{\theta_{pxk}}^{\pi/6} p_x^2 J_{px} \csc \beta_{pxpx} \exp[j(\theta_{Jpx} + 2\theta_{px} + \beta_{pxpx})] d\theta_{px} \quad (260)$$

$$\bar{J}_y = \int_{-\infty}^{\infty} \bar{J}_{py} d\bar{p}_y \quad (261)$$

$$= \int_{-\infty}^{\infty} J_{py} \sec \beta_{pypy} \exp[j(\theta_{Jpy} + \theta_{py} + \beta_{pypy})] dp_y \quad (262)$$

$$= \int_0^{\pi/6} p_y J_{py} \csc \beta_{pypy} \exp[j(\theta_{Jpy} + \theta_{py} + \beta_{pypy})] d\theta_{py} \quad (263)$$

$$\bar{J}_z = \int_{-\infty}^{\infty} \bar{J}_{pz} d\bar{p}_z \quad (264)$$

$$= \int_{-\infty}^{\infty} J_{pz} \sec \beta_{pzpz} \exp[j(\theta_{Jpz} + \theta_{pz} + \beta_{pzpz})] dp_z \quad (265)$$

$$= \int_0^{\pi/6} p_z J_{pz} \csc \beta_{pzpz} \exp[j(\theta_{Jpz} + \theta_{pz} + \beta_{pzpz})] d\theta_{pz} \quad (266)$$

where

$$\bar{J}_{px} = J_{px} \exp(j\theta_{Jpx}) = \exp[-(\bar{p}_x^2 - 2m_e \bar{u})/(2m_e kT)] \quad (267)$$

$$\bar{J}_{py} = J_{py} \exp(j\theta_{Jpy}) = \exp[-\bar{p}_y^2/(2m_e kT)] \quad (268)$$

$$\bar{J}_{pz} = J_{pz} \exp(j\theta_{Jpz}) = \exp[-\bar{p}_z^2/(2m_e kT)] \quad (269)$$

$$J_{px} = \exp\{-[p_x^2 \cos(2\theta_{px}) - 2m_e \mu \cos \theta_\mu]/(2m_e kT)\} \quad (270)$$

$$J_{py} = \exp\{-[p_y^2 \cos(2\theta_{py})]/(2m_e kT)\} \quad (271)$$

$$J_{pz} = \exp\{-[p_z^2 \cos(2\theta_{pz})]/(2m_e kT)\} \quad (272)$$

$$\theta_{Jpx} = -[p_x^2 \sin(2\theta_{px}) - 2m_e \mu \sin \theta_\mu]/(2m_e kT) \quad (273)$$

$$\theta_{Jpy} = -[p_y^2 \sin(2\theta_{py})]/(2m_e kT) \quad (274)$$

$$\theta_{Jpz} = -[p_z^2 \sin(2\theta_{pz})]/(2m_e kT) \quad (275)$$

The lower limits of integration  $\bar{p}_{xk}$ ,  $p_{xk}$  and  $\theta_{pxk}$  in the integrals of equations (258) through (260) are given by equations (237), (245) and (246). The upper limits of the integrals over the phase angles of the momenta in equations (260), (263) and (266) are obtained from the assumption of spacetime coherence which gives<sup>13,14,20</sup>

$$\theta_{px}^c = \theta_x^c - \theta_t^c = \pi/3 - \pi/6 = \pi/6 \quad (276)$$

with similar expressions for the y and z directions.

The quantities  $J_x$ ,  $\theta_{Jx}$ ,  $J_y$ ,  $\theta_{Jy}$ ,  $J_z$  and  $\theta_{Jz}$  that are required for the determination of the measured thermionic emission current given in equation (257) can be obtained by first calculating the real and imaginary parts of the integrals given in equations (259), (260), (262), (263), (265) and (266) as follows

$$J_x \cos \theta_{Jx} = \int_{p_{xk}}^{\infty} p_x J_{px} \sec \beta_{pxpx} \cos(\theta_{Jpx} + 2\theta_{px} + \beta_{pxpx}) dp_x \quad (277)$$

$$= \int_{\theta_{pxk}}^{\pi/6} p_x^2 J_{px} \csc \beta_{pxpx} \cos(\theta_{Jpx} + 2\theta_{px} + \beta_{pxpx}) d\theta_{px} \quad (278)$$

$$J_x \sin \theta_{Jx} = \int_{p_{xk}}^{\infty} p_x J_{px} \sec \beta_{pxpx} \sin(\theta_{Jpx} + 2\theta_{px} + \beta_{pxpx}) dp_x \quad (279)$$

$$= \int_{\theta_{pxk}}^{\pi/6} p_x^2 J_{px} \csc \beta_{pxpx} \sin(\theta_{Jpx} + 2\theta_{px} + \beta_{pxpx}) d\theta_{px} \quad (280)$$

$$J_y \cos \theta_{Jy} = \int_{-\infty}^{\infty} J_{py} \sec \beta_{pypy} \cos(\theta_{Jpy} + \theta_{py} + \beta_{pypy}) dp_y \quad (281)$$

$$= \int_0^{\pi/6} p_y J_{py} \csc \beta_{pypy} \cos(\theta_{Jpy} + \theta_{py} + \beta_{pypy}) d\theta_{py} \quad (282)$$

$$J_y \sin \theta_{Jy} = \int_{-\infty}^{\infty} J_{py} \sec \beta_{pypy} \sin(\theta_{Jy} + \theta_{py} + \beta_{pypy}) dp_y \quad (283)$$

$$= \int_0^{\pi/6} p_y J_{py} \csc \beta_{pypy} \sin(\theta_{Jy} + \theta_{py} + \beta_{pypy}) d\theta_{py} \quad (284)$$

$$J_z \cos \theta_{Jz} = \int_{-\infty}^{\infty} J_{pz} \sec \beta_{pzpz} \cos(\theta_{Jpz} + \theta_{pz} + \beta_{pzpz}) dp_z \quad (285)$$

$$= \int_0^{\pi/6} p_z J_{pz} \csc \beta_{pzpz} \cos(\theta_{Jpz} + \theta_{pz} + \beta_{pzpz}) d\theta_{pz} \quad (286)$$

$$J_z \sin \theta_{Jz} = \int_{-\infty}^{\infty} J_{pz} \sec \beta_{pzpz} \sin(\theta_{Jpz} + \theta_{pz} + \beta_{pzpz}) dp_z \quad (287)$$

$$= \int_0^{\pi/6} p_z J_{pz} \csc \beta_{pzpz} \sin(\theta_{Jpz} + \theta_{pz} + \beta_{pzpz}) d\theta_{pz} \quad (288)$$

These integrals will be evaluated for special cases in the following sections.

### C. Thermionic Emission from the Normal State of a High- $T_c$ Material.

Consider the case of the normal state of a high- $T_c$  material with  $T > T_c$  for which the electrons are in a nearly incoherent spacetime state which is described by

$$\beta_{xx} \sim 0 \quad \beta_{yy} \sim 0 \quad \beta_{zz} \sim 0 \quad (289A)$$

$$\beta_{tt}^x \sim 0 \quad \beta_{tt}^y \sim 0 \quad \beta_{tt}^z \sim 0 \quad (289B)$$

$$\beta_{pxpx} \sim 0 \quad \beta_{pypy} \sim 0 \quad \beta_{pzpz} \sim 0 \quad (290)$$

which corresponds to

$$\theta_x^i = \text{constant} \quad \theta_y^i = \text{constant} \quad \theta_z^i = \text{constant} \quad (291)$$

$$\theta_t^{xi} = \text{constant} \quad \theta_t^{yi} = \text{constant} \quad \theta_t^{zi} = \text{constant} \quad (292)$$

From equations (23), (289), (291) and (292) it follows that

$$\theta_{px}^i = \theta_x^i - \theta_t^{xi} = \text{constant} \quad (293)$$

$$\theta_{py}^i = \theta_y^i - \theta_t^{yi} = \text{constant} \quad (294)$$

$$\theta_{pz}^i = \theta_z^i - \theta_t^{zi} = \text{constant} \quad (295)$$

Then the integrals in equations (277), (279), (281) and (283) become

$$J_x^i \cos \theta_{Jx}^i = W_{1x} \cos(2\theta_{px}^i) + W_{2x} \sin(2\theta_{px}^i) \quad (296)$$

$$J_x^i \sin \theta_{Jx}^i = W_{1x} \sin(2\theta_{px}^i) - W_{2x} \cos(2\theta_{px}^i) \quad (297)$$

$$J_y^i \cos \theta_{Jy}^i = U_{1y} \cos \theta_{py}^i + U_{2y} \sin \theta_{py}^i \quad (298)$$

$$J_y^i \sin \theta_{Jy}^i = U_{1y} \sin \theta_{py}^i - U_{2y} \cos \theta_{py}^i \quad (299)$$

$$J_z^i \cos \theta_{Jz}^i = U_{1z} \cos \theta_{pz}^i + U_{2z} \sin \theta_{pz}^i \quad (300)$$

$$J_z^i \sin \theta_{Jz}^i = U_{1z} \sin \theta_{pz}^i - U_{2z} \cos \theta_{pz}^i \quad (301)$$

where

$$W_{1x} = \int_{p_{xk}}^{\infty} p_x J_{px}^i \cos \theta_{Jpx}^i dp_x \quad (302)$$

$$W_{2x} = - \int_{p_{xk}}^{\infty} p_x J_{px}^i \sin \theta_{Jpx}^i dp_x \quad (303)$$

$$U_{1y} = \int_{-\infty}^{\infty} J_{py}^i \cos \theta_{Jpy}^i dp_y \quad (304)$$

$$U_{2y} = - \int_{-\infty}^{\infty} J_{py}^i \sin \theta_{Jpy}^i dp_y \quad (305)$$

$$U_{1z} = \int_{-\infty}^{\infty} J_{pz}^i \cos \theta_{Jpz}^i dp_z \quad (306)$$

$$U_{2z} = - \int_{-\infty}^{\infty} J_{pz}^i \sin \theta_{Jpz}^i dp_z \quad (307)$$

where

$$J_{px}^i = \exp(g - c_x p_x^2) \quad (308)$$

$$J_{py}^i = \exp(-c_y p_y^2) \quad (309)$$

$$J_{pz}^i = \exp(-c_z p_z^2) \quad (310)$$



$$\theta_{J_{px}}^i = a - b_x p_x^2 \quad (311)$$

$$\theta_{J_{py}}^i = -b_y p_y^2 \quad (312)$$

$$\theta_{J_{pz}}^i = -b_z p_z^2 \quad (313)$$

and where

$$c_x = \cos(2\theta_{px}^i)/(2m_e kT) \quad (314)$$

$$c_y = \cos(2\theta_{py}^i)/(2m_e kT) \quad (315)$$

$$c_z = \cos(2\theta_{pz}^i)/(2m_e kT) \quad (316)$$

$$b_x = \sin(2\theta_{px}^i)/(2m_e kT) \quad (317)$$

$$b_y = \sin(2\theta_{py}^i)/(2m_e kT) \quad (318)$$

$$b_z = \sin(2\theta_{pz}^i)/(2m_e kT) \quad (319)$$

$$g = (\mu \cos \theta_\mu)/(kT) \quad (320)$$

$$a = (\mu \sin \theta_\mu)/(kT) \quad (321)$$

The integrals in equations (302) through (307) will now be evaluated.

Combining equations (302) through (321) allows the integrals in equations (302) through (307) to be written as

$$W_{1x} = \int_{p_{xk}}^{\infty} \exp(g - c_x p_x^2) \cos(a - b_x p_x^2) p_x dp_x \quad (322)$$

$$W_{2x} = - \int_{p_{xk}}^{\infty} \exp(g - c_x p_x^2) \sin(a - b_x p_x^2) p_x dp_x \quad (323)$$

$$U_{1y} = \int_{-\infty}^{\infty} \exp(-c_y p_y^2) \cos(b_y p_y^2) dp_y \quad (324)$$

$$U_{2y} = \int_{-\infty}^{\infty} \exp(-c_y p_y^2) \sin(b_y p_y^2) dp_y \quad (325)$$

$$U_{1z} = \int_{-\infty}^{\infty} \exp(-c_z p_z^2) \cos(b_z p_z^2) dp_z \quad (326)$$

$$U_{2z} = \int_{-\infty}^{\infty} \exp(-c_z p_z^2) \sin(b_z p_z^2) dp_z \quad (327)$$

The integrals in equations (322) and (323) can be written as

$$W_{1x} = B(U_{1x} \cos a + U_{2x} \sin a) \quad (328)$$

$$W_{2x} = B(U_{2x} \cos a - U_{1x} \sin a) \quad (329)$$

where

$$U_{1x} = \int_{p_{xk}}^{\infty} \exp(-c_x p_x^2) \cos(b_x p_x^2) p_x dp_x \quad (330)$$

$$U_{2x} = \int_{p_{xk}}^{\infty} \exp(-c_x p_x^2) \sin(b_x p_x^2) p_x dp_x \quad (331)$$

where

$$B = e^g \quad (332)$$

where  $g$  is given by equation (320).

The evaluation of the integrals  $U_{1v}$  and  $U_{2v}$  for  $v = x, y, z$  can be obtained from integral tables.<sup>26</sup> The results are as follows

$$U_{1x} = m_e kT \exp(-c_x p_{xk}^2) \cos(2\theta_{px}^i + b_x p_{xk}^2) \quad (333)$$

$$U_{2x} = m_e kT \exp(-c_x p_{xk}^2) \sin(2\theta_{px}^i + b_x p_{xk}^2) \quad (334)$$

$$W_{1x} = m_e kT \exp(g - c_x p_{xk}^2) \cos(2\theta_{px}^i + b_x p_{xk}^2 - a) \quad (335)$$

$$W_{2x} = m_e kT \exp(g - c_x p_{xk}^2) \sin(2\theta_{px}^i + b_x p_{xk}^2 - a) \quad (336)$$

$$U_{1y} = (2\pi m_e kT)^{1/2} \cos \theta_{py}^i \quad (337)$$

$$U_{2y} = (2\pi m_e kT)^{1/2} \sin \theta_{py}^i \quad (338)$$

$$U_{1z} = (2\pi m_e kT)^{1/2} \cos \theta_{pz}^i \quad (339)$$

$$U_{2z} = (2\pi m_e kT)^{1/2} \sin \theta_{pz}^i \quad (340)$$

If all internal phase angles are set equal to zero these integrals become

$$U_{1x}^0 = m_e kT \exp[-p_{xk}^2 / (2m_e kT)] \quad (341)$$

$$U_{2x}^0 = 0 \quad (342)$$

$$W_{1x}^0 = m_e kT \exp\{[\mu - p_{xk}^2/(2m_e)]/(kT)\} \quad (343)$$

$$W_{2x}^0 = 0 \quad (344)$$

$$U_{1y}^0 = (2\pi m_e kT)^{1/2} \quad (345)$$

$$U_{2y}^0 = 0 \quad (346)$$

$$U_{1z}^0 = (2\pi m_e kT)^{1/2} \quad (347)$$

$$U_{2z}^0 = 0 \quad (348)$$

The quantities  $U_{1v}$ ,  $U_{2v}$ ,  $W_{1x}$  and  $W_{2x}$  enter directly into the calculation of the thermionic emission current.

The calculation of the factors  $J_v^i$  and their associated internal phase angles  $\theta_{Jv}^i$  for  $v = x, y, z$  that appear in equations (255) through (257) for the thermionic emission current from the normal state follows from equations (296) through (301) using the results presented in equations (333) through (340) as

$$J_x^i \cos \theta_{Jx}^i = m_e kT \exp(g - c_x p_{xk}^2) \cos(a - b_x p_{xk}^2) \quad (340)$$

$$J_x^i \sin \theta_{Jx}^i = m_e kT \exp(g - c_x p_{xk}^2) \sin(a - b_x p_{xk}^2) \quad (341)$$

$$J_y^i \cos \theta_{Jy}^i = (2\pi m_e kT)^{1/2} \quad (351)$$

$$J_y^i \sin \theta_{Jy}^i = 0 \quad (352)$$

$$J_z^i \cos \theta_{Jz}^i = (2\pi m_e kT)^{1/2} \quad (353)$$

$$J_z^i \sin \theta_{Jz}^i = 0 \quad (354)$$

Equations (349) through (354) give

$$J_x^i = m_e kT \exp(g - c_x p_{xk}^2) \quad (355)$$

$$\theta_{Jx}^i = a - b_x p_{xk}^2 \quad (356)$$

$$J_y^i = (2\pi m_e kT)^{1/2} \quad (357)$$

$$\theta_{Jy}^i = 0 \quad (358)$$

$$J_z^1 = (2\pi m_e kT)^{1/2} \quad (359)$$

$$\theta_{Jz}^1 = 0 \quad (360)$$

Then the thermionic emission current for the normal state of a high- $T_c$  material is obtained from equations (257) and (355) through (360) to be

$$I_{xm}^1 = A_o T^2 \exp(g - c_x p_{xk}^2) \cos(a - b_x p_{xk}^2) \quad (361)$$

where  $A_o$  is given by equation (233), and  $g, a, c_x, b_x$  and  $p_{xk}$  are given by equations (320), (321), (314), (317) and (246) respectively. When all internal phase angles are set equal to zero equation (361) reduces to the Richardson-Dushman equation (232) because for this case

$$g^o = \mu/(kT) \quad (362)$$

$$c_x^o = (2m_e kT)^{-1} \quad (363)$$

$$p_{xk}^o = [2m_e(\mu + e\phi)]^{1/2} \quad (364)$$

$$a^o = 0 \quad (365)$$

$$b_x^o = 0 \quad (366)$$

$$g^o - c_x^o p_{xk}^{o2} = -e\phi/(kT) \quad (367)$$

which corresponds to incoherent spacetime.

The predicted measured value of the thermionic emission current for the normal state of a high- $T_c$  material is given by equation (361) within the approximation of constant internal phase angles for the single particle momenta, i.e.,  $\theta_{px} = \theta_{px}^1$ ,  $\theta_{py} = \theta_{py}^1$  and  $\theta_{pz} = \theta_{pz}^1$ . The predicted measured value of the thermionic emission current is seen to have a leading temperature term of  $T^2$  that is augmented by a temperature dependent exponential and trigonometrical term. For the case of an ordinary metal where all internal phase angles are set equal to zero equation (361) reduces to the standard Richardson-Dushman equation (232).

#### D. Thermionic Emission from the Superconducting State of a High- $T_c$ Material.

The following conditions are valid for the coherent spacetime of the superconducting state of a high- $T_c$  material

$$\beta_{xx}^c = \pi/2 \quad \beta_{yy}^c = \pi/2 \quad \beta_{zz}^c = \pi/2 \quad (368A)$$

$$\beta_{tt}^{xc} = \pi/2 \quad \beta_{tt}^{yc} = \pi/2 \quad \beta_{tt}^{zc} = \pi/2 \quad (368B)$$

It is further assumed that equations (106) and (107) are valid for the superconducting state of a high- $T_c$  material. Then according to equation (108) the following coherence conditions are valid for incipient unbound electron pairs from which come the electrons of thermionic emission from the superconducting state

$$\beta_{pxpx}^{c+} = \pi/2 \quad \beta_{pypy}^{c+} = \pi/2 \quad \beta_{pzip}^{c+} = \pi/2 \quad (369)$$

so that equation (115) gives with  $\theta_{px}$ ,  $\theta_{py}$  and  $\theta_{pz}$  as variables

$$d\bar{p}_x^{c+} = j\bar{p}_x^c d\theta_{px} = jp_{xc} d\theta_{px} \exp(j\theta_{px}) \quad (370)$$

$$d\bar{p}_y^{c+} = j\bar{p}_y^c d\theta_{py} = jp_{yc} d\theta_{py} \exp(j\theta_{py}) \quad (371)$$

$$d\bar{p}_z^{c+} = j\bar{p}_z^c d\theta_{pz} = jp_{zc} d\theta_{pz} \exp(j\theta_{pz}) \quad (372)$$

and where

$$p_{xc}^2 = 2m_e kT_{cx} \quad (373)$$

$$p_{yc}^2 = 2m_e kT_{cy} \quad (374)$$

$$p_{zc}^2 = 2m_e kT_{cz} \quad (375)$$

where

$T_{cx}$  = superconducting transition temperature for ac plane

$T_{cy}$  = superconducting transition temperature for bc plane

$T_{cz}$  = superconducting transition temperature for ab plane

For a bulk superconductor the superconducting transition temperature is given by

$$p_{xc}^2 = p_{yc}^2 = p_{zc}^2 = 2m_e kT_c \quad (376)$$

$$T_{cx} = T_{cy} = T_{cz} = T_c \quad (377)$$

where  $T_c$  is the common value of the superconductivity transition temperature.

The complex number thermionic emission current is given by equations (247), (260), (263) and (266) for the superconducting state of a high- $T_c$  material, where for this case

$$\bar{J}_x^c = jp_{xc}^2 \int_{\theta_{pxk}}^{\pi/6} J_{px} \exp[j(\theta_{Jpx} + 2\theta_{px})] d\theta_{px} \quad (378)$$

$$\bar{J}_y^c = j p_{yc} \int_0^{\pi/6} J_{py} \exp[j(\theta_{Jpy} + \theta_{py})] d\theta_{py} \quad (379)$$

$$\bar{J}_z^c = j p_{zc} \int_0^{\pi/6} J_{pz} \exp[j(\theta_{Jpz} + \theta_{pz})] d\theta_{pz} \quad (380)$$

where  $J_{px}$ ,  $J_{py}$ ,  $J_{pz}$ ,  $\theta_{Jpx}$ ,  $\theta_{Jpy}$  and  $\theta_{Jpz}$  are given by equations (270) through (275). The real and imaginary parts of equations (378) through (380) are given by

$$J_x^c \cos \theta_{Jx}^c = - p_{xc}^2 \int_{\theta_{pxk}}^{\pi/6} J_{px} \sin(\theta_{Jpx} + 2\theta_{px}) d\theta_{px} \quad (381)$$

$$J_x^c \sin \theta_{Jx}^c = p_{xc}^2 \int_{\theta_{pxk}}^{\pi/6} J_{px} \cos(\theta_{Jpx} + 2\theta_{px}) d\theta_{px} \quad (382)$$

$$J_y^c \cos \theta_{Jy}^c = - p_{yc} \int_0^{\pi/6} J_{py} \sin(\theta_{Jpy} + \theta_{py}) d\theta_{py} \quad (383)$$

$$J_y^c \sin \theta_{Jy}^c = p_{yc} \int_0^{\pi/6} J_{py} \cos(\theta_{Jpy} + \theta_{py}) d\theta_{py} \quad (384)$$

$$J_z^c \cos \theta_{Jz}^c = - p_{zc} \int_0^{\pi/6} J_{pz} \sin(\theta_{Jpz} + \theta_{pz}) d\theta_{pz} \quad (385)$$

$$J_z^c \sin \theta_{Jz}^c = p_{zc} \int_0^{\pi/6} J_{pz} \cos(\theta_{Jpz} + \theta_{pz}) d\theta_{pz} \quad (386)$$

The problem is to determine  $J_x^c$ ,  $J_y^c$ ,  $J_z^c$ ,  $\theta_{Jx}^c$ ,  $\theta_{Jy}^c$  and  $\theta_{Jz}^c$ .

Combining equations (270) through (275) with equations (381) through (386) gives

$$J_x^c \cos \theta_{Jx}^c = p_{xc}^2 e^g (K_1 \cos a - K_2 \sin a) \quad (387)$$

$$J_x^c \sin \theta_{Jx}^c = p_{xc}^2 e^g (K_1 \sin a + K_2 \cos a) \quad (388)$$

$$J_y^c \cos \theta_{Jy}^c = p_{yc} K_3 \quad (389)$$

$$J_y^c \sin \theta_{Jy}^c = p_{yc} K_4 \quad (390)$$

$$J_z^c \cos \theta_{Jz}^c = p_{zc} K_5 \quad (391)$$

$$J_z^c \sin \theta_{Jz}^c = p_{zc} K_6 \quad (392)$$

where  $g$  and  $a$  are given by equations (320) and (321), and where

$$K_1 = \int_{\theta_{pxk}}^{\pi/6} \exp[-\alpha_{xc} \cos(2\theta_{px})] \sin[\alpha_{xc} \sin(2\theta_{px}) - 2\theta_{px}] d\theta_{px} \quad (393)$$

$$K_2 = \int_{\theta_{pxk}}^{\pi/6} \exp[-\alpha_{xc} \cos(2\theta_{px})] \cos[\alpha_{xc} \sin(2\theta_{px}) - 2\theta_{px}] d\theta_{px} \quad (394)$$

$$K_3 = \int_0^{\pi/6} \exp[-\alpha_{yc} \cos(2\theta_{py})] \sin[\alpha_{yc} \sin(2\theta_{py}) - \theta_{py}] d\theta_{py} \quad (395)$$

$$K_4 = \int_0^{\pi/6} \exp[-\alpha_{yc} \cos(2\theta_{py})] \cos[\alpha_{yc} \sin(2\theta_{py}) - \theta_{py}] d\theta_{py} \quad (396)$$

$$K_5 = \int_0^{\pi/6} \exp[-\alpha_{zc} \cos(2\theta_{pz})] \sin[\alpha_{zc} \sin(2\theta_{pz}) - \theta_{pz}] d\theta_{pz} \quad (397)$$

$$K_6 = \int_0^{\pi/6} \exp[-\alpha_{zc} \cos(2\theta_{pz})] \cos[\alpha_{zc} \sin(2\theta_{pz}) - \theta_{pz}] d\theta_{pz} \quad (398)$$

where

$$\alpha_{xc} = p_{xc}^2 / (2m_e kT) = T_{cx} / T \quad (399)$$

$$\alpha_{yc} = p_{yc}^2 / (2m_e kT) = T_{cy} / T \quad (400)$$

$$\alpha_{zc} = p_{zc}^2 / (2m_e kT) = T_{cz} / T \quad (401)$$

Equations (387) through (391) give

$$J_x^c = p_{xc}^2 e^g (K_1^2 + K_2^2)^{1/2} \quad (401)$$

$$J_y^c = p_{yc}^2 (K_3^2 + K_4^2)^{1/2} \quad (402)$$

$$J_z^c = p_{zc}^2 (K_5^2 + K_6^2)^{1/2} \quad (403)$$

$$\tan \theta_{Jx}^c = (K_1 \sin a + K_2 \cos a) / (K_1 \cos a - K_2 \sin a) \quad (405)$$

$$\tan \theta_{Jy}^c = K_4 / K_3 \quad (406)$$

$$\tan \theta_{Jz}^c = K_6 / K_5 \quad (407)$$

Equations (402) through (407) can be used to calculate the thermoemission from the superconducting state of a high- $T_c$  compound.

The measured thermionic emission current for a high- $T_c$  superconductor is given by equation (257). Several cases can be considered according to whether the spacetime coordinates are coherent or incoherent. In all cases considered the thermionic emission is along the x axis.

Case a. Spacetime Coherence Along All Axis.

Equations (257), (373) through (375) and (402) through (407) give the measured thermionic emission current as

$$I_{xm}^{ccc} = 2e/(m_e h^3) J_x^c J_y^c J_z^c \cos(\theta_{Jx}^c + \theta_{Jy}^c + \theta_{Jz}^c) \quad (408)$$

or equivalently

$$I_{xm}^{ccc} = 2/\pi A_o T_{cx} (T_{cy} T_{cz})^{1/2} e^g G_{123456} \cos \phi_{xyz}^{ccc} \quad (409)$$

where  $A_o$  is the Richardson-Dushman constant given by equation (233);  $T_{cx}$ ,  $T_{cy}$  and  $T_{cz}$  are the superconductivity transition temperatures,  $g$  is given by equation (320), and where

$$G_{123456} = [(K_1^2 + K_2^2)(K_3^2 + K_4^2)(K_5^2 + K_6^2)]^{1/2} \quad (410)$$

$$\phi_{xyz}^{ccc} = \theta_{Jx}^c + \theta_{Jy}^c + \theta_{Jz}^c \quad (411)$$

where  $K_1$  through  $K_6$  are given by equations (393) through (398), and where  $\theta_{Jx}^c$ ,  $\theta_{Jy}^c$  and  $\theta_{Jz}^c$  are given by equations (405) through (407). Equation (409) gives the measured thermionic emission current for the superconducting state and is valid for

$$T < T_{cx} \quad T < T_{cy} \quad T < T_{cz} \quad (412)$$

For the case of a bulk superconductor (bs) equation (377) is valid and equation (409) becomes

$$I_{xm}^{bs} = 2/\pi A_o T_c^2 e^g G_{123456}^{bs} \cos \phi_{xyz}^{bs} \quad (413)$$

where now

$$K_3 = K_5 \quad K_4 = K_6 \quad \theta_{Jy}^c = \theta_{Jz}^c \quad (414)$$

$$G_{123456}^{bs} = (K_1^2 + K_2^2)^{1/2} (K_3^2 + K_4^2) \quad (415)$$

$$\phi_{xyz}^{bs} = \theta_{Jx}^c + 2 \theta_{Jy}^c \quad (416)$$



which is valid for  $T < T_c$ . The leading temperature terms in equations (409) or (413) are functions of the superconducting transition temperatures and are not related to the  $T^2$  term that appears in the Richardson-Dushman equation that describes thermionic emission from ordinary metals. The temperature dependence of the thermionic emission current comes essentially through the functions  $e^g$  and  $K_1$  through  $K_6$ . Equations (320), (393) through (398), (409) and (410) show that

$$e^g G_{123456} \rightarrow 0 \quad T \rightarrow 0 \quad (417)$$

so that there is no thermionic emission at  $T = 0$

$$I_{xm}^{ccc}(T = 0) = 0 \quad (418)$$

The functions  $K_n(T)$  increase slowly with temperature from their values of  $K_n(0) = 0$  at  $T = 0$ .

Case b. Spacetime Coherence Along the Emission Axis, and Spacetime Coherence and Incoherence Along the Two Perpendicular Axes.

For this case equation (257) is written as

$$I_{xm}^{cci} = 2e/(m_e h^3) J_x^c J_y^c J_z^i \cos(\theta_{Jx}^c + \theta_{Jy}^c + \theta_{Jz}^i) \quad (419)$$

where  $J_x^c$ ,  $J_y^c$  and  $J_z^i$  are given by equations (402), (403) and (353) while  $\theta_{Jx}^c$ ,  $\theta_{Jy}^c$  and  $\theta_{Jz}^i$  are given by equations (405), (406) and (360). Equation (419) can be rewritten as

$$I_{xm}^{cci} = 2/\sqrt{\pi} A_o T_{cx}^{1/2} T_{cy}^{1/2} e^g G_{1234} \cos \phi_{xyz}^{cci} \quad (420)$$

where

$$G_{1234} = [(K_1^2 + K_2^2)(K_3^2 + K_4^2)]^{1/2} \quad (421)$$

$$\begin{aligned} \phi_{xyz}^{cci} &= \theta_{Jx}^c + \theta_{Jy}^c + \theta_{Jz}^i \\ &= \theta_{Jx}^c + \theta_{Jy}^c \end{aligned} \quad (422)$$

For the case when  $T_{cx} = T_{cy} = T_c$  equation (420) becomes

$$I_{xm}^{cci} = 2/\sqrt{\pi} A_o T_c^{3/2} e^g G_{1234} \cos \phi_{xyz}^{cci} \quad (423)$$

If the incoherent spacetime axis is taken to be the y axis the thermionic emission current is given analogously to equations (419) and (420) as

$$I_{xm}^{cic} = 2e/(m_e h^3) J_x^c J_y^i J_z^c \cos(\theta_{Jx}^c + \theta_{Jy}^i + \theta_{Jz}^c) \quad (424)$$

$$= 2/\sqrt{\pi} A_o T_{cx} T_{cz}^{1/2} T^{1/2} e^g G_{1256} \cos \phi_{xyz}^{cic} \quad (425)$$

where

$$G_{1256} = [(K_1^2 + K_2^2)(K_5^2 + K_6^2)]^{1/2} \quad (426)$$

$$\begin{aligned} \phi_{xyz}^{cic} &= \theta_{Jx}^c + \theta_{Jy}^i + \theta_{Jz}^c \\ &= \theta_{Jx}^c + \theta_{Jz}^c \end{aligned} \quad (427)$$

Again, for the case  $T_{cx} = T_{cz} = T_c$  equation (425) becomes

$$I_{xm}^{cic} = 2/\sqrt{\pi} A_o T_c^{3/2} T^{1/2} e^g G_{1256} \cos \phi_{xyz}^{cic} \quad (428)$$

which is essentially equivalent to equation (423). Note the leading  $T^{1/2}$  behavior of the thermionic emission current in equations (420), (423), (425) and (428) which is different from the  $T^2$  behavior associated with the Richardson-Dushman equation for ordinary metals. The  $T^{1/2}$  behavior may possibly be experimentally verified in the high- $T_c$  compounds. Equations (420) and (425) are valid for  $T < T_{cx}$ ,  $T < T_{cy}$  and  $T < T_{cz}$ , while equations (423) and (428) are valid for  $T < T_c$ .

Case c. Spacetime Coherence Along the Emission Axis, and Spacetime Incoherence Along Both Perpendicular Axes.

In this case equation (257) becomes

$$I_{xm}^{cii} = 2e/(m_e h^3) J_x^c J_y^i J_z^i \cos(\theta_{Jx}^c + \theta_{Jy}^i + \theta_{Jz}^i) \quad (429)$$

where  $J_x^c$ ,  $J_y^i$  and  $J_z^i$  are given by equations (402), (357) and (359), and where  $\theta_{Jx}^c$ ,  $\theta_{Jy}^i$  and  $\theta_{Jz}^i$  are given by equations (405), (358) and (360). Equation (429) can be rewritten as

$$I_{xm}^{cii} = 2A_o T_c T e^g G_{12} \cos \phi_{xyz}^{cii} \quad (430)$$

where  $T_{cx} = T_c$  and where

$$G_{12} = (K_1^2 + K_2^2)^{1/2} \quad (431)$$

$$\begin{aligned} \phi_{xyz}^{cii} &= \theta_{Jx}^c + \theta_{Jy}^i + \theta_{Jz}^i \\ &= \theta_{Jx}^c \end{aligned} \quad (432)$$

For this case the leading temperature dependent term is  $T$  which possibly could be experimentally verified in special classes of high- $T_c$  materials. Equation (430) is valid for  $T < T_c$ .

Case d. Incoherent Spacetime Along the Emission Axis, and Spacetime Coherence Along Both Perpendicular Axes.

For this case equation (257) is written as

$$I_{xm}^{icc} = 2e/(m_e h^3) J_x^i J_y^c J_z^c \cos(\theta_{Jx}^i + \theta_{Jy}^c + \theta_{Jz}^c) \quad (433)$$

where  $J_x^i$ ,  $J_y^c$  and  $J_z^c$  are given by equations (355), (403) and (404) and  $\theta_{Jx}^i$ ,  $\theta_{Jy}^c$  and  $\theta_{Jz}^c$  are given by equations (356), (406) and (407). Then equation (433) can be rewritten as

$$I_{xm}^{icc} = A_o / \pi T_{cy}^{1/2} T_{cz}^{1/2} T e^{g'} G_{3456} \exp \phi_{xyz}^{icc} \quad (434)$$

where

$$g' = g - c_x p_{xc}^2 \quad (435)$$

$$G_{3456} = [(K_3^2 + K_4^2)(K_5^2 + K_6^2)]^{1/2} \quad (436)$$

$$\phi_{xyz}^{icc} = \theta_{Jx}^i + \theta_{Jy}^c + \theta_{Jz}^c \quad (437)$$

If  $T_{cy} = T_{cz} = T_c$  then equation (434) simplifies to

$$I_{xm}^{icc} = A_o / \pi T_c T e^{g'} G_{3456} \cos \phi_{xyz}^{icc} \quad (438)$$

Equation (434) is valid for  $T < T_{cy}$  and  $T < T_{cz}$ , while equation (438) is valid for  $T < T_c$ . Note the linear dependence of the leading temperature dependent terms in equations (434) and (438).

Case e. Incoherent Spacetime Along Emission Axis, and Spacetime Coherence and Incoherence Along the Two Perpendicular Axes.

In this case equation (257) is written as

$$I_{xm}^{iic} = 2e/(m_e h^3) J_x^i J_y^i J_z^c \cos(\theta_{Jx}^i + \theta_{Jy}^i + \theta_{Jz}^c) \quad (439)$$

where  $J_x^i$ ,  $J_y^i$  and  $J_z^c$  are given by equations (355), (357) and (404), while  $\theta_{Jx}^i$ ,  $\theta_{Jy}^i$  and  $\theta_{Jz}^c$  are given by equations (356), (358) and (407). Equation (439) can be rewritten as

$$I_{xm}^{iic} = A_o / \sqrt{\pi} T_{cz}^{1/2} T^{3/2} e^{g'} G_{56} \cos \phi_{xyz}^{iic} \quad (440)$$

where  $g'$  is given by equation (435) and where

$$G_{56} = (K_5^2 + K_6^2)^{1/2} \quad (441)$$

$$\begin{aligned} \phi_{xyz}^{iic} &= \theta_{Jx}^i + \theta_{Jy}^i + \theta_{Jz}^c \\ &= \theta_{Jx}^i + \theta_{Jz}^c \end{aligned} \quad (442)$$

For this case the leading temperature term of the thermionic emission current is  $T^{3/2}$ . If the coherent spacetime axis is selected to be the y axis the thermionic emission current is given by

$$I_{xm}^{ici} = 2e/(m_e h^3) J_x^i J_y^c J_z^i \cos(\theta_{Jx}^i + \theta_{Jy}^c + \theta_{Jz}^i) \quad (443)$$

$$= A_0 / \sqrt{\pi} T_{cy}^{1/2} T^{3/2} e^{g'} G_{34} \cos \phi_{xyz}^{ici} \quad (444)$$

where

$$G_{34} = (K_3^2 + K_4^2)^{1/2} \quad (445)$$

$$\begin{aligned} \phi_{xyz}^{ici} &= \theta_{Jx}^i + \theta_{Jy}^c + \theta_{Jz}^i \\ &= \theta_{Jx}^i + \theta_{Jy}^c \end{aligned} \quad (446)$$

For this case the leading temperature dependent term is  $T^{3/2}$ .

From equations (409), (413), (420), (425), (430), (434), (438), (440), and (444) it is clear that, depending on whether spacetime is coherent or incoherent along the emission axis and along the two perpendicular axes, the leading temperature terms of the thermionic emission current for the superconducting state of a high- $T_c$  material can be either  $T^0$ ,  $T^{1/2}$ ,  $T$  or  $T^{3/2}$ .

**4. CONCLUSION.** The superconducting state ( $T < T_c$ ) of a high- $T_c$  material is associated with coherent spacetime while the normal state ( $T > T_c$ ) is related to partially coherent spacetime. For the superconducting state the Cooper electrons rotate incoherently or coherently but their radial motion in space and time is in a coherent state. The motion of the electrons under a weak attractive inverse square force is non-Keplerian because the rotational period is independent of the separation distance of the two electrons in a Cooper pair. The thermionic emission currents for the normal and superconducting states of a high- $T_c$  compound have been calculated. For the normal state of a high- $T_c$  material the thermionic emission current has a Richardson-Dushman form that is modified by an exponential and cosine term, but for the superconducting state the thermionic emission current has a completely different form and does not have the leading quadratic temperature dependence of the Richardson-Dushman equation but rather goes as  $T^0$ ,  $T^{1/2}$ ,  $T$  or  $T^{3/2}$  depending on whether spacetime is coherent or incoherent along the thermionic emission axis and along the two perpendicular axes.

## ACKNOWLEDGEMENT

I wish to thank Elizabeth K. Klein for typing and editing this paper.

## REFERENCES

1. Bednorz, J. G. and Müller, K. A., editors, Earlier and Recent Aspects of Superconductivity, Springer-Verlag, New York, 1990.
2. Burns, G., High-Temperature Superconductivity, Academic, New York, 1992.
3. Phillips, J. C., Physics of High- $T_c$  Superconductors, Academic, New York, 1989.
4. Lynn, J. W., editor, High Temperature Superconductivity, Springer-Verlag, New York, 1990.
5. Ishiguro, T. and Yamaji, K., Organic Superconductors, Springer-Verlag, New York, 1990.
6. Goodstein, D. L., States of Matter, Dover, New York, 1975.
7. Isihara, A., Condensed Matter Physics, Oxford, New York, 1991.
8. Tinkham, M., Introduction to Superconductivity, Krieger, Malabar, Florida, 1980.
9. Tilley, D. R. and Tilley, J., Superfluidity and Superconductivity, Adam Hilger, New York, 1990.
10. Edwards, H. L., Markert, J. T. and Lozanne, A. L., "Energy Gap and Surface Structure of  $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$  Probed by Scanning Tunneling Microscopy," *Phys. Rev. Lett.*, Vol. 69, p. 2967, 16 November 1992.
11. Degiorgi, L., Wachter, P., Huang, S., Wiley, J. and Kaner, R., "Optical Response of the Superconducting State of  $\text{K}_3\text{C}_{60}$  and  $\text{RbC}_{60}$ ," *Phys. Rev. Lett.*, Vol. 69, p. 2987, 16 November 1992.
12. Zhang, Z., Chen, C. and Lieber, C., "Tunneling Spectroscopy of  $\text{M}_3\text{C}_{60}$  Superconductors: The Energy Gap, Strong Coupling, and Superconductivity," *Science*, Vol. 254, p. 1619, 13 December 1991.
13. Weiss, R. A., "Electromagnetism and Gravity," paper in Clean Fission, K&W Publications, Vicksburg, MS, 1992.
14. Weiss, R. A., "High- $T_c$  Superconductivity and the Photoelectric Effect," paper in Clean Fission, K&W Publications, Vicksburg, MS, 1992.
15. Margaritondo, G., "100 Years of Photoemission," *Physics Today*, p. 66, April 1988.
16. Vaterlaus, A., Milani, F. and Meier, F., "Spin Polarization of Thermoemitted Electrons from Cesium Ni and Fe," *Phys. Rev. Lett.*, Vol. 65, 10 December 1990.

17. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
18. Weiss, R. A., Clean Fission, K&W Publications, Vicksburg, MS, 1992.
19. Weiss, R. A., Relativistic Thermodynamics, Vols. 1&2, K&W Publications, Vicksburg, MS, 1976.
20. Weiss, R. A., "Dynamical Systems in Asymmetric Space and Time," article in Clean Fission, K&W Publications, Vicksburg, MS, 1992.
21. Goldstein, H., Classical Mechanics, Addison-Wesley, New York, 1980.
22. Powell, J. L. and Crasemann, B., Quantum Mechanics, Addison-Wesley, Reading, 1961.
23. Born, M., Atomic Physics, Hafner, New York, 1935.
24. Eyring, H., Henderson, D., Stover, B. J. and Eyring, E. M., Statistical Mechanics and Dynamics, John Wiley, New York, 1964.
25. Fowler, R. H., Statistical Mechanics, Cambridge Univ. Press, New York, 1955.
26. Gradshteyn, I. S. and Ryzhik, I. M., Table of Integrals, Series, and Products, Academic, New York, 1980.

## SLOW AND ULTRAFAST DYNAMICAL SYSTEMS

Richard A. Weiss

U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** Newton's law of dynamics is written in a complex number form that is valid for space and time coordinates that exhibit broken internal symmetries. The space and time coordinates and therefore also the potential function are written as complex numbers in an internal space. This allows the possibility of motion in space and time where the magnitudes of the complex number space and time coordinates change while the phase angles remain constant, and the possibility of internal space motions where the phase angles of the space and time coordinates vary while the magnitudes of the space and time coordinates are constants which corresponds to the case of internal rotations in spacetime. Similarly the magnitude of a complex number potential can change while the internal phase angle of the potential remains fixed as in the case a a slow mechanical process, or the internal phase angle of the complex number potential can change for a fixed magnitude of the potential which is the case of an ultrafast mechanical process. This leads to eight possible forms for Newton's dynamical law of motion corresponding to slow and ultrafast mechanical processes in coherent or incoherent space and for coherent and incoherent time.

**1. INTRODUCTION.** Classical mechanics is formulated against a background of space and time which is represented as an inert continuous medium. Bodies in motion are represented as trajectories in the space and time background. In this context space and time is a mathematical construct that itself has no reality other than geometry because space and time coordinates can be chosen in an arbitrary manner. However, the development of quantum field theory has suggested that the vacuum is a real physical medium having well defined properties.<sup>1-3</sup> The vacuum state influences the matter within the vacuum through such effects a vacuum polarization and the self energy of fundamental particles.<sup>1-3</sup>

The reality of the vacuum state suggests that space and time coordinates may be more than just mathematical constructs, and it has been suggested that space and time coordinates are complex numbers in an internal space.<sup>4</sup> The nature of the complex number coordinates has its origin in the relativistic trace equation which implies that pressure is a complex number in an internal space.<sup>4,5</sup> The internal phase angles of the space and time coordinates are affected by the presence of external fields such as gravity and electromagnetism.<sup>4</sup> The concept of complex number spacetime allows the possibility of coherent motion in spacetime wherein the complex number coordinates rotate in an internal space with fixed magnitudes of the space and time coordinates and where the internal phase angles of the space and time coordinates are now the dynamical variables of motion. Expressions for the velocity and acceleration of particles undergoing internal spacetime motions have already appeared in the literature.<sup>4,6</sup> This paper introduces the concepts of slow and ultrafast mechanical processes. A slow mechanical process occurs when the magnitude of the complex number potential function varies while the phase angle of the potential remains constant. The case of an ultrafast mechanical process exists when the complex number potential function rotates in an internal space with the magnitude of the potential held

fixed while the phase angle of the potential changes.

The space and time coordinates must be written as complex numbers in internal space as follows<sup>4,6</sup>

$$\bar{v} = v \exp(j\theta_v) \quad (1)$$

$$\bar{t} = t \exp(j\theta_t^v) \quad (2)$$

where  $v = x, y, z$  for cartesian coordinates;  $r, \phi, z$  for cylindrical polar coordinates; and  $\rho, \phi, \psi$  for spherical polar coordinates. Strictly speaking the internal phase angle of time  $\theta_t^v$  is associated with each space coordinate so that for the three elementary coordinate systems the internal phase angles of time are:

$$\theta_t^x, \theta_t^y, \theta_t^z \quad \theta_t^r, \theta_t^\phi, \theta_t^z \quad \theta_t^\rho, \theta_t^\phi, \theta_t^\psi \quad (3)$$

The following quantities often appear in the calculations involving partially coherent broken symmetry space and time<sup>4,6</sup>

$$\tan \beta_{vv} = v \partial \theta_v / \partial v \quad (4)$$

$$\tan \beta_{tt}^v = t \partial \theta_t^v / \partial t \quad (5)$$

as for example in the following differentials<sup>4,6</sup>

$$\begin{aligned} d\bar{v} &= \sec \beta_{vv} dv \exp[j(\theta_v + \beta_{vv})] \\ &= \csc \beta_{vv} v d\theta_v \exp[j(\theta_v + \beta_{vv})] \end{aligned} \quad (6)$$

$$\begin{aligned} d\bar{t} &= \sec \beta_{tt}^v dt \exp[j(\theta_t^v + \beta_{tt}^v)] \\ &= \csc \beta_{tt}^v t d\theta_t^v \exp[j(\theta_t^v + \beta_{tt}^v)] \end{aligned} \quad (7)$$

For incoherent spacetime  $\beta_{vv} = 0$  and  $\beta_{tt}^v = 0$  while for coherent spacetime  $\beta_{vv} = \pi/2$  and  $\beta_{tt}^v = \pi/2$ . The measured coordinates of space and time are given by<sup>4,6</sup>

$$v_m = v \cos \theta_v \quad t_m^v = t \cos \theta_t^v \quad (8)$$

The complex number coordinate speed is obtained from equations (6) and (7) as

$$\bar{v}_v = v_v \exp(j\theta_{vv}) = d\bar{v}/d\bar{t} \quad (9)$$

where



$$v_v = \cos \beta_{tt}^v \sec \beta_{vv} dv/dt \quad (10)$$

$$= \cos \beta_{tt}^v \csc \beta_{vv} v d\theta_v/dt \quad (11)$$

$$= \sin \beta_{tt}^v \sec \beta_{vv} t^{-1} dv/d\theta_t^v \quad (12)$$

$$= \sin \beta_{tt}^v \csc \beta_{vv} v/t d\theta_v/d\theta_t^v \quad (13)$$

where

$$\theta_{vv} = \theta_v + \beta_{vv} - \theta_t^v - \beta_{tt}^v \quad (14)$$

and where  $v = x, y, z; r, \phi, z; \text{ or } \rho, \phi, \psi$ . The particle velocity, momentum and energy are written as complex numbers in internal space as follows<sup>4,6</sup>

$$\bar{v}_v = v_v \exp(j\theta_{vv}) \quad \bar{p}_v = p_v \exp(j\theta_{pv}) \quad \bar{E} = E \exp(j\theta_E) \quad (15)$$

and the corresponding measured quantities are

$$v_{vm} = v_v \cos \theta_{vv} \quad p_{vm} = p_v \cos \theta_{pv} \quad E_m = E \cos \theta_E \quad (16)$$

which are the real values of the complex number quantities.

The basic interest in the realization of complex number spacetime by the application of external fields is the possibility of internal motions in spacetime which may perhaps eventually be used for developing power sources such as rocket engines. The essential outline of this paper is as follows: Section 2 outlines the subject of kinematics in spacetime with broken internal symmetries, and Section 3 considers Newton's law of motion for slow and ultrafast mechanical processes in asymmetric spacetime.

**2. KINEMATICS IN SPACE AND TIME WITH BROKEN INTERNAL SYMMETRIES.** This section develops the basic expressions for particle speed and acceleration in spacetime with broken internal symmetries. The concept of motion in totally coherent spacetime is introduced. The connections are made between the measured kinematical quantities of partially coherent spacetime and the conventionally calculated kinematical quantities of incoherent spacetime.

#### A. Particle Speed.

The speed of a particle in broken symmetry spacetime is obtained from equations (9) through (14) as<sup>4,6</sup>

$$\bar{v}_x = v_x \exp(j\theta_{vx}) = d\bar{x}/d\bar{t} \quad (17)$$

where

$$v_x = \cos \beta_{tt}^x \sec \beta_{xx} dx/dt \quad (18)$$

$$= \cos \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt \quad (19)$$

$$= \sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x \quad (20)$$

$$= \sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x \quad (21)$$

and

$$\theta_{vx} = \theta_x + \beta_{xx} - \theta_t^x - \theta_{tt}^x \quad (22)$$

where  $\beta_{xx}$  and  $\beta_{tt}^x$  are given by equations (4) and (5) respectively. The measured particle speed is given by equations (16) and (18) through (21) as<sup>6</sup>

$$v_{mx} = v_x \cos \theta_{vx} \quad (23)$$

The single particle momentum is then given by

$$\bar{p}_x = p_x \exp(j\theta_{px}) \quad p_x = mv_x \quad \theta_{px} = \theta_{vx} \quad (24)$$

Equations similar to (17) through (22) can be developed for the y and z coordinates.

The special forms of the particle speed for the four possible spacetime states will now be presented.

Case a. Incoherent Space and Incoherent Time.

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \theta_t = 0 \quad \beta_{tt}^x = 0 \quad (25)$$

Combining equations (18) and (25) gives

$$v_x^{ii} = dx/dt \quad \theta_{vx}^{ii} = 0 \quad (26)$$

which is the conventional description.

Case b. Coherent Space and Incoherent Time.

$$\beta_{xx} = \pi/2 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (27)$$

Combining equations (19) and (27) gives

$$v_x^{ci} = x d\theta_x/dt \quad \theta_{vx}^{ci} = \theta_x + \pi/2 \quad (28)$$

which describe internal space motion in external time.

Case c. Incoherent Space and Coherent Time.

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \beta_{tt}^x = \pi/2 \quad (29)$$

Combining equations (20) and (29) gives

$$v_x^{ic} = t^{-1} dx/d\theta_t^x \quad \theta_{vx}^{ic} = -\theta_t^x - \pi/2 \quad (30)$$

which describes an internal time motion in external space.

Case d. Coherent Space and Coherent Time.

$$\beta_{xx} = \pi/2 \quad \beta_{tt}^x = \pi/2 \quad (31)$$

Combining equations (21) and (31) gives

$$v_x^c = x/t \, d\theta_x/d\theta_t^x \quad \theta_{vx}^c = \theta_x - \theta_t^x \quad (32)$$

which describes internal motion in both space and time.

B. Particle Acceleration.

For a broken symmetry spacetime the particle acceleration is given by equations (17) through (22) as<sup>4,6</sup>

$$\bar{a}_x = a_x \exp(j\theta_{ax}) = d\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2 \quad (33)$$

where<sup>6</sup>

$$a_x = \sec \beta_{vxvx} \cos \beta_{tt}^x \, dv_x/dt \quad (34)$$

$$= \csc \beta_{vxvx} \cos \beta_{tt}^x \, v_x \, d\theta_{vx}/dt \quad (35)$$

$$= \sec \beta_{vxvx} \sin \beta_{tt}^x \, t^{-1} \, dv_x/d\theta_t^x \quad (36)$$

$$= \csc \beta_{vxvx} \sin \beta_{tt}^x \, v_x/t \, d\theta_{vx}/d\theta_t^x \quad (37)$$

and<sup>6</sup>

$$\theta_{ax} = \theta_{vx} + \beta_{vxvx} - \theta_t^x - \beta_{tt}^x \quad (38)$$

where<sup>6</sup>

$$\tan \beta_{vxvx} = v_x \partial \theta_{vx} / \partial v_x \quad (39)$$

From Newton's law of motion written as

$$\bar{F}_x = F_x \exp(j\theta_{Fx}) = m\bar{a}_x \quad (40)$$

it follows that

$$F_x = ma_x \quad \theta_{Fx} = \theta_{ax} \quad (41)$$

The measured acceleration is given by

$$a_{mx} = a_x \cos \theta_{ax} \quad (42)$$

The measured force is given by equations (40) and (41) as

$$F_{mx} = F_x \cos \theta_{Fx} \quad (43)$$

Combining equations (41) through (43) gives

$$F_{mx} = ma_{mx} \quad (44)$$

and therefore the measured acceleration is determined from the measured force by Newton's law of motion. The values of the acceleration magnitude  $a_x$  and acceleration internal phase angle  $\theta_{ax}$  will now be calculated for several cases of the broken symmetry of space and time.

Case a. Incoherent Space and Incoherent Time.

A general expression for the acceleration is developed that can be used to deduce the limiting case of incoherent space and incoherent time which is described by

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (45)$$

The appropriate expressions for the acceleration magnitude and internal phase angle are deduced from equations (34) and (38) to be<sup>6</sup>

$$a_x = \sec \beta_{v_x v_x} \cos \beta_{tt}^x dv_x/dt \quad (46)$$

$$\theta_{ax} = \theta_{vx} + \beta_{v_x v_x} - \theta_t^x - \beta_{tt}^x \quad (47)$$

where

$$\tan \beta_{v_x v_x} = v_x \partial \theta_{vx} / \partial v_x \quad (48)$$

Combining equations (18) and (46) gives<sup>6</sup>

$$a_x = \cos \beta_{tt}^x \sec \beta_{v_x v_x} d/dt(\cos \beta_{tt}^x \sec \beta_{xx} dx/dt) \quad (49)$$

An alternative expression for the acceleration is obtained from equations (18)

and (35) as

$$\begin{aligned}
 a_x &= \csc \beta_{vxvx} \cos \beta_{tt}^x v_x d\theta_{vx}/dt \\
 &= \csc \beta_{vxvx} \cos \beta_{tt}^x (\cos \beta_{tt}^x \sec \beta_{xx} dx/dt) d\theta_{vx}/dt \\
 &= \csc \beta_{vxvx} \cos^2 \beta_{tt}^x \sec \beta_{xx} dx/dt d\theta_{vx}/dt
 \end{aligned} \tag{50}$$

Combining equations (22) and (47) gives the phase angle of the acceleration as

$$\theta_{ax} = \theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) \tag{51}$$

When the conditions in equation (45) are valid the expression in equation (49) reduces to the standard case of incoherent space and incoherent time.

Case b. Coherent Space and Incoherent Time.

General expressions for the acceleration are now deduced which can be used to make the transition to the case of coherent space and incoherent time which is described by

$$\beta_{xx} = \pi/2 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \tag{52}$$

An appropriate general expression for the acceleration magnitude and internal phase angle for this case is obtained from equations (19) and (34) to be<sup>6</sup>

$$\begin{aligned}
 a_x &= \sec \beta_{vxvx} \cos \beta_{tt}^x dv_x/dt \\
 &= \sec \beta_{vxvx} \cos \beta_{tt}^x d/dt(\cos \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt)
 \end{aligned} \tag{53}$$

An alternative form of the acceleration is obtained from equation (35) as

$$a_x = \csc \beta_{vxvx} \cos \beta_{tt}^x v_x d\theta_{vx}/dt \tag{54}$$

where  $v_x$  is given by equation (19). Combining equations (19) and (54) gives

$$a_x = \csc \beta_{vxvx} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \tag{55}$$

where from equation (22)

$$d\theta_{vx}/dt = d/dt(\theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x) \tag{56}$$

The acceleration phase angle is obtained from equation (51) as

$$\theta_{ax} = \theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) \tag{57}$$

For the case at hand it is convenient to write the acceleration in equation (33) for negative acceleration (attractive forces) as

$$\begin{aligned}\bar{a}_x &= a_x \exp(j\theta_{ax}) = a_x^\dagger \exp(j\theta_{ax}^\dagger) \\ &= d\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2\end{aligned}\quad (58)$$

where

$$a_x^\dagger = -a_x \quad (59)$$

$$\theta_{ax}^\dagger = \theta_{ax} - \pi \quad (60)$$

so that

$$a_x^\dagger = -\csc \beta_{vxvx} \cos \beta_{tt}^x v_x d\theta_{vx}/dt \quad (61)$$

$$= -\csc \beta_{vxvx} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \quad (62)$$

$$\theta_{ax}^\dagger = \theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) - \pi \quad (63)$$

which is an equivalent description of the acceleration.

For the special case of coherent space and incoherent time, equations (52), (56), (62) and (63) give

$$a_x^{ci\dagger} = -\csc \beta_{vxvx}^{ci} x(d\theta_x/dt)^2 \quad (64)$$

$$\theta_{ax}^{ci\dagger} = \theta_x + \beta_{vxvx}^{ci} - \pi/2 \quad (65)$$

where from equations (19), (22), (48) and (52) it follows that

$$v_x^{ci} = x d\theta_x/dt \quad (66)$$

$$\theta_{vx}^{ci} = \theta_x + \pi/2 \quad (67)$$

$$\tan \beta_{vxvx}^{ci} = E_{xt}^{ci}/F_{xt}^{ci} \quad (68)$$

where

$$E_{xt}^{ci} = (d\theta_x/dt)^2 \quad E_{xt}^{ci} \geq 0 \quad (69)$$

$$F_{xt}^{ci} = d^2\theta_x/dt^2 \quad F_{xt}^{ci} \leq 0 \quad (70)$$

Equation (68) then gives

$$\csc \beta_{vxvx}^{ci} = (E_{xt}^{ci})^{-1} [(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} \quad (71)$$

Because  $F_{xt}^{ci} \leq 0$  it follows from equation (68) that

$$\beta_{vxvx}^{ci} = \pi/2 + \kappa_{xt} \quad (72)$$

where  $\kappa_{xt} \geq 0$  is a small number which is also given by

$$\tan \kappa_{xt} = |F_{xt}^{ci}|/E_{xt}^{ci} \quad (73)$$

Combining equations (64), (65), (71) and (72) gives

$$a_x^{ci+} = -x[(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} \quad (74)$$

$$\theta_{ax}^{ci+} = \theta_x + \kappa_{xt} \quad (75)$$

The angle  $\theta_{ax}^{ci+}$  is generally a small number. Clearly the acceleration in this case is directed opposite to the displacement  $x$ . Finally, from equations (58) and (64) through (75) it follows that for a negative acceleration (attractive forces)

$$\bar{a}_x^{ci} = -\bar{x}(E_{xt}^{ci} - jF_{xt}^{ci}) \quad (76)$$

The positive acceleration of a particle moving under the influence of a repulsive force in coherent space and coherent time is given by

$$\bar{a}_x^{ci+} = \bar{x}(E_{xt}^{ci+} - jF_{xt}^{ci+}) \quad (77)$$

$$a_x^{ci+} = x[(E_{xt}^{ci+})^2 + (F_{xt}^{ci+})^2]^{1/2} \quad (78)$$

$$\theta_{ax}^{ci+} = \theta_x + \beta_{vxvx}^{ci+} - \pi/2 \quad (79)$$

with

$$E_{xt}^{ci+} = E_{xt}^{ci} \quad E_{xt}^{ci+} \geq 0 \quad (80)$$

$$F_{xt}^{ci+} = F_{xt}^{ci} \quad F_{xt}^{ci+} \leq 0 \quad (81)$$

$$\tan \beta_{vxvx}^{ci+} = E_{xt}^{ci+}/F_{xt}^{ci+} \quad (82)$$

$$\beta_{vxvx}^{ci+} = \beta_{vxvx}^{ci} = \pi/2 + \kappa_{xt} \quad (83)$$

so that

$$\theta_{ax}^{ci+} = \theta_{ax}^{ci+} = \theta_x + \kappa_{xt} \quad a_x^{ci+} = -a_x^{ci+} \quad (84)$$

$$\bar{a}_x^{ci+} = -\bar{a}_x^{ci} = \bar{x}(E_{xt}^{ci} - jF_{xt}^{ci}) \quad (85)$$

Case c. Incoherent Space and Coherent Time.

An expression for the acceleration is now derived from which the limiting case of incoherent space and coherent time can be obtained. This limiting case is described by

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \beta_{tt}^x = \pi/2 \quad (86)$$

The required general expression for the acceleration magnitude and acceleration phase angle is obtained from equations (36) and (51) as<sup>6</sup>

$$a_x = \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x \quad (87)$$

$$\theta_{ax} = \theta_x + \beta_{vxvx} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) \quad (88)$$

Combining equations (20) and (87) gives

$$a_x = \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \quad (89)$$

From equations (20) and (37) it follows that

$$a_x = \csc \beta_{vxvx} \sin \beta_{tt}^x v_x/t d\theta_{vx}/d\theta_t^x \quad (90)$$

$$= \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} dx/d\theta_t^x d\theta_{vx}/d\theta_t^x \quad (91)$$

For this case it is convenient to rewrite the acceleration equation (33) in the following form

$$\bar{a}_x = a_x \exp(j\theta_{ax}) = a'_x \exp(j\theta'_{ax}) = d\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2 \quad (92)$$

where

$$a'_x = -a_x \quad (93)$$

$$\theta'_{ax} = \theta_{ax} + \pi \quad (94)$$

so that an equivalent representation of the acceleration that is useful for attractive forces and negative accelerations is obtained from equations (93) and (94) and (87) through (89) as

$$a'_x = -\sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x \quad (95)$$

$$= -\sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \quad (96)$$



$$\theta'_{ax} = \theta_{ax} + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) + \pi \quad (97)$$

where  $x$  and  $\theta_t^x$  are variables.

In the case of incoherent space and coherent time, equations (86), (96) and (97) give

$$a_x^{ic'} = - \sec \beta_{vxvx}^{ic} t^{-2} d^2x/d\theta_t^{x2} \quad (98)$$

$$\theta_{ax}^{ic'} = \beta_{vxvx}^{ic} - 2\theta_t^x \quad (99)$$

where  $\theta_{ax}^{ic'}$  is a small number, and where from equations (20), (22), (48) and (86) it follows that

$$v_x^{ic} = t^{-1} dx/d\theta_t^x \quad (100)$$

$$\theta_{vx}^{ic} = -\theta_t^x - \pi/2 \quad (101)$$

$$\tan \beta_{vxvx}^{ic} = E_{xt}^{ic}/F_{xt}^{ic} \quad (102)$$

where

$$E_{xt}^{ic} = -dx/d\theta_t^x \quad E_{xt}^{ic} \geq 0 \quad (103)$$

$$F_{xt}^{ic} = d^2x/d\theta_t^{x2} \quad F_{xt}^{ic} \geq 0 \quad (104)$$

$$\sec \beta_{vxvx}^{ic} = (F_{xt}^{ic})^{-1} [(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} \quad (105)$$

Because  $F_{xt}^{ic} \geq 0$  and  $F_{xt}^{ic} \geq 0$  it follows that  $\beta_{vxvx}^{ic}$  is a small positive number for this case. Equations (98) and (105) give

$$a_x^{ic'} = -t^{-2} [(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} \quad (106)$$

Finally from equations (92) and (98) through (106) it follows that for an attractive force with a negative acceleration

$$\bar{a}_x^{ic} = -1/t^2 (F_{xt}^{ic} + jE_{xt}^{ic}) \quad (107)$$

For the case of a positive acceleration due to a repulsive force acting in incoherent space and coherent time the acceleration is given by

$$\bar{a}_x^{ic+} = 1/t^2 (F_{xt}^{ic+} + jE_{xt}^{ic}) \quad (108)$$

$$a_x^{ic+} = t^{-2} [(E_{xt}^{ic+})^2 + (F_{xt}^{ic+})^2]^{1/2} \quad (109)$$

$$\theta_{ax}^{ic+} = \beta_{vxvx}^{ic+} - 2\theta_t^x \quad (110)$$

with

$$E_{xt}^{ic+} = E_{xt}^{ic} \quad E_{xt}^{ic+} \geq 0 \quad (111)$$

$$F_{xt}^{ic+} = F_{xt}^{ic} \quad F_{xt}^{ic+} \geq 0 \quad (112)$$

$$\tan \beta_{vxvx}^{ic+} = E_{xt}^{ic+} / F_{xt}^{ic+} \quad (113)$$

$$\beta_{vxvx}^{ic+} = \beta_{vxvx}^{ic} \quad (114)$$

so that for positive accelerations

$$\theta_{ax}^{ic+} = \theta_{ax}^{ic'} = \beta_{vxvx}^{ic} - 2\theta_t^x \quad a_x^{ic+} = -a_x^{ic'} \quad (115)$$

$$\bar{a}_x^{ic+} = -\bar{a}_x^{ic} = 1/t^2 (F_{xt}^{ic} + jE_{xt}^{ic}) \quad (116)$$

Case d. Coherent Space and Coherent Time.

In this section an equation for the acceleration of a particle is obtained which can be used to attain the limit of coherent space and coherent time which is defined by

$$\beta_{xx} = \pi/2 \quad \beta_{tt}^x = \pi/2 \quad (117)$$

The general expression for the magnitude and internal phase angle of the acceleration is obtained from equations (37) and (51) as<sup>6</sup>

$$a_x = \csc \beta_{vxvx} \sin \beta_{tt}^x v_x/t \, d\theta_{vx}/d\theta_t^x \quad (118)$$

$$\theta_{ax} = \theta_x + \beta_{vxvx} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) \quad (119)$$

Combining equations (21) and (118) gives the acceleration magnitude as

$$a_x = \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} x/t^2 \, d\theta_x/d\theta_t^x \, d\theta_{vx}/d\theta_t^x \quad (120)$$

where from equation (22) it follows that

$$d\theta_{vx}/d\theta_t^x = d\theta_x/d\theta_t^x - 1 + d/d\theta_t^x (\beta_{xx} - \beta_{tt}^x) \quad (121)$$

Another expression for the acceleration magnitude can be obtained from equations (21) and (36) which gives

$$a_x = \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} \, d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t \, d\theta_x/d\theta_t^x) \quad (122)$$

In the present case it is convenient to write the acceleration in equation (33) as

$$\bar{a}_x = a_x \exp(j\theta_{ax}) = a'_x \exp(j\theta'_{ax}) = d\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2 \quad (123)$$

where

$$a'_x = -a_x \quad (124)$$

$$\theta'_{ax} = \theta_{ax} + \pi \quad (125)$$

and an alternative representation of the acceleration that is suitable for attractive forces is

$$a'_x = -\csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} x/t^2 d\theta_x/d\theta_t^x d\theta_{vx}/d\theta_t^x \quad (126)$$

$$= -\sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \quad (127)$$

$$\theta'_{ax} = \theta_{vx} + \beta_{vxvx} - \theta_t^x - \beta_{tt}^x + \pi \quad (128)$$

$$= \theta_x + \beta_{vxvx} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) + \pi$$

For the case of coherent space and coherent time equations (126) through (128) become with the help of equation (117)

$$a'_x = -\csc \beta_{vxvx}^c x/t^2 d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad (129)$$

$$= -\sec \beta_{vxvx}^c x/t^2 d^2\theta_x/d\theta_t^{x2} \quad (130)$$

$$\theta'_{ax} = \theta_x + \beta_{vxvx}^c - 2\theta_t^x + \pi/2 \quad (131)$$

From equations (21) and (22) it follows that

$$v_x^c = x/t d\theta_x/d\theta_t^x \quad (132)$$

$$\theta_{vx}^c = \theta_x - \theta_t^x \quad (133)$$

Equations (48), (132) and (133) give

$$\tan \beta_{vxvx}^c = E_{xt}^c / F_{xt}^c \quad (134)$$

where for coherent spacetime

$$E_{xt}^c = d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad E_{xt}^c \leq 0 \quad (135)$$

$$F_{xt}^c = d^2\theta_x/d\theta_t^{x2} \quad F_{xt}^c \geq 0 \quad (136)$$

$$\csc \beta_{vxvx}^c = (E_{xt}^c)^{-1} [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} \quad (137)$$

$$\sec \beta_{vxvx}^c = (F_{xt}^c)^{-1} [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} \quad (138)$$

where  $d\theta_x/d\theta_t^x \leq 1$ . In a gravitational field, for example, the following relationship holds<sup>4</sup>

$$2\theta_t^x \sim 3\theta_x \quad d\theta_x/d\theta_t^x \sim 2/3 \quad (139)$$

Therefore in general  $E_{xt}^c \leq 0$  and  $F_{xt}^c \geq 0$ , and it is convenient to introduce a new angle by writing

$$\beta_{vxvx}^c = -\pi/2 + \delta_{xt} \quad (140)$$

or equivalently

$$\tan \delta_{xt} = F_{xt}^c / |E_{xt}^c| \quad (141)$$

so that in general  $\delta_{xt} \geq 0$ . Combining equations (129) through (131) and (135) through (138) gives for coherent space and time

$$a_x^{c'} = -x/t^2 [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} \quad (142)$$

$$\theta_{ax}^{c'} = \theta_{vx}^c + \beta_{vxvx}^c - \theta_t^x + \pi/2 \quad (143)$$

$$= \theta_{vx}^c - \theta_t^x + \delta_{xt}$$

$$= \theta_x + \beta_{vxvx}^c - 2\theta_t^x + \pi/2$$

$$= \theta_x - 2\theta_t^x + \delta_{xt}$$

Therefore  $\theta_{ax}^{c'}$  is a small number. Coherent space and time represents an internal motion in space and time that can be written in complex number form as

$$d\bar{x} = j\bar{x}d\theta_x \quad d\bar{t} = j\bar{t}d\theta_t^x \quad (144)$$

which is equivalent to equation (117). The magnitude and phase angle equations (132) and (133) are equivalent to the following complex number expression for the particle speed in coherent spacetime

$$\bar{v}_x^c = (d\bar{x}/d\bar{t})^c = \bar{x}/\bar{t} d\theta_x/d\theta_t^x \quad (145)$$

The magnitude and phase angle of the acceleration that are given in equations (142) and (143) correspond to a complex number acceleration that is obtained from equation (123) and is given by

$$\begin{aligned}\bar{a}_x^c &= (d^2\bar{x}/d\bar{t}^2)^c = \bar{x}/\bar{t}^2 [d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) - jd^2\theta_x/d\theta_t^{x2}] \\ &= -\bar{x}/\bar{t}^2 (-E_{xt}^c + jF_{xt}^c)\end{aligned}\quad (146)$$

as it must be because of the definition of  $\delta_{xt}$  given in equations (140) and (141).

For a repulsive force acting on a particle in coherent space and coherent time the positive acceleration is given by

$$\bar{a}_x^{c+} = \bar{x}/\bar{t}^2 (-E_{xt}^{c+} + jF_{xt}^{c+}) \quad (147)$$

$$a_x^{c+} = x/t^2 [(E_{xt}^{c+})^2 + (F_{xt}^{c+})^2]^{1/2} \quad (148)$$

$$\theta_{ax}^{c+} = \theta_x + \beta_{vxvx}^{c+} - 2\theta_t^x + \pi/2 \quad (149)$$

with

$$E_{xt}^{c+} = E_{xt}^c \quad E_{xt}^{c+} \leq 0 \quad (150)$$

$$F_{xt}^{c+} = F_{xt}^c \quad F_{xt}^{c+} \geq 0 \quad (151)$$

$$\tan \beta_{vxvx}^{c+} = E_{xt}^{c+}/F_{xt}^{c+} \quad (152)$$

$$\beta_{vxvx}^{c+} = \beta_{vxvx}^c = -\pi/2 + \delta_{xt} \quad (153)$$

and

$$\theta_{ax}^{c+} = \theta_{ax}^{c'} = \theta_x - 2\theta_t^x + \delta_{xt} \quad a_x^{c+} = -a_x^{c'} \quad (154)$$

$$\bar{a}_x^{c+} = -\bar{a}_x^c = \bar{x}/\bar{t}^2 (-E_{xt}^c + jF_{xt}^c) \quad (155)$$

which agrees with equation (146).

3. SLOW AND ULTRAFast FORMS OF NEWTON'S LAW OF MOTION IN BROKEN SYMMETRY SPACETIME. Newton's law of motion for a particle in a potential field can be written for spacetime with broken internal symmetries as follows<sup>7-10</sup>

$$m\bar{a}_x = md^2\bar{x}/d\bar{t}^2 = -\partial\bar{W}/\partial\bar{x} = \bar{F}_x \quad (156)$$

where  $m$  = particle mass and  $\bar{W}$  = complex number potential which can be written as

$$\bar{W} = W \exp(j\theta_W) \quad (157)$$

The derivatives of the potential function can be written in four ways. Representing the complex number force in the following way

$$\bar{F}_x = F_x \exp(j\theta_{Fx}) = -\partial\bar{W}/\partial\bar{x} \quad (158)$$

gives<sup>6</sup>

$$F_x = -\cos \beta_{xx} \sec \beta_{ww} \partial W / \partial x \quad (159)$$

$$= -\cos \beta_{xx} \csc \beta_{ww} W \partial \theta_w / \partial x \quad (160)$$

$$= -\sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W / \partial \theta_x \quad (161)$$

$$= -\sin \beta_{xx} \csc \beta_{ww} W/x \partial \theta_w / \partial \theta_x \quad (162)$$

and

$$\theta_{Fx} = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} \quad (163)$$

where  $\theta_{Fx}$  is a small angle, and where

$$\tan \beta_{ww} = W \partial \theta_w / \partial W \quad (164)$$

For repulsive forces with  $F_x > 0$  the phase angle and magnitude equations equivalent to equation (156) are  $\theta_{ax} = \theta_{Fx}$  and  $ma_x = F_x$ . For attractive forces the phase angle and magnitude equations that are equivalent to Newton's law in equation (156) for nearly incoherent space and nearly incoherent time are obtained from equations (51) and (163) for  $F_x < 0$  and  $a_x < 0$  as

$$\theta_{ax} = \theta_{Fx} \quad ma_x = F_x \quad (165)$$

For nearly coherent space and nearly incoherent time, equations (59), (60) and (163) give for  $F_x < 0$  and  $a_x > 0$

$$\theta_{ax}^\dagger = \theta_{Fx} \quad -ma_x = F_x \quad (166)$$

For nearly incoherent space and nearly coherent time, equations (93), (94) and (163) give for  $F_x < 0$  and  $a_x > 0$

$$\theta_{ax}' = \theta_{Fx} \quad -ma_x = F_x \quad (167)$$

For the case of nearly coherent space and nearly coherent time it follows from equations (124), (125) and (163) that for  $F_x < 0$  and  $a_x > 0$

$$\theta_{ax}' = \theta_{Fx} \quad -ma_x = F_x \quad (168)$$

Newton's law of motion given by equation (156) will now be considered for the four kinematic spacetime conditions that were considered in Section 2 and for slow and ultrafast mechanical processes, which produces eight possible cases for Newton's law of motion.

Case 1. Slow Process, Incoherent Space and Incoherent Time.

This section develops the Newtonian law of dynamics for broken symmetry spacetime in a form that is suitable to make the transition to the case of a slow mechanical process in incoherent space and incoherent time which is described by

$$\theta_W = 0 \quad \beta_{WW} = 0 \quad (169)$$

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (170)$$

where  $x$  and  $t$  are variables. Combining equations (40), (46), (47), (156) (159) and (163) gives Newton's law as

$$m \cos \beta_{tt}^x \sec \beta_{v xv x} dv_x/dt = - \cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \quad (171)$$

$$\begin{aligned} \theta_{ax} &= \theta_{vx} + \beta_{v xv x} - \theta_t^x - \beta_{tt}^x \\ &= \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \end{aligned} \quad (172)$$

where  $\beta_{v xv x}$  and  $\beta_{WW}$  are given by equations (48) and (164) respectively. Combining equations (18) and (171) gives Newton's law of motion as

$$\begin{aligned} m \cos \beta_{tt}^x \sec \beta_{v xv x} d/dt(\cos \beta_{tt}^x \sec \beta_{xx} dx/dt) \\ = - \cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \end{aligned} \quad (173)$$

Combining equations (50) and (159) gives an alternative form of Newton's law

$$\begin{aligned} m \csc \beta_{v xv x} \cos^2 \beta_{tt}^x \sec \beta_{xx} dx/dt d\theta_{vx}/dt \\ = - \cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \end{aligned} \quad (174)$$

Combining equations (51) and (172) gives

$$\begin{aligned} \theta_{ax} &= \theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) \\ &= \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \end{aligned} \quad (175)$$

If the potential function is given by a power law  $W \sim x^{-\sigma}$  then equation (164) gives

$$\beta_{WW} = \beta_{xx} \quad (176)$$

Therefore a reasonable approximation to equation (175) for many potential functions is given by

$$2\theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) \sim \theta_W \quad (177)$$

The limiting case of a slow mechanical process in incoherent space and incoherent time is obtained by setting all phase angles equal to zero and the exact equation (173) becomes

$$m d^2x/dt^2 = - \partial W/\partial x \quad (178)$$

which is the standard form of Newton's law of motion.

#### Case 2. Slow Process, Coherent Space and Incoherent Time.

This section gives a form of Newton's law of motion that can be used to regain the special case of a slow dynamical process in coherent space and incoherent time which is described by

$$\beta_{WW} = 0 \quad \theta_W = 0 \quad (179)$$

$$\beta_{xx} = \pi/2 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (180)$$

where  $\theta_x$  and  $t$  are variables. Combining equations (19), (53), (59) and (161) gives Newton's law of motion for attractive forces as

$$- m \sec \beta_{v xv x} \cos \beta_{tt}^x dv_x/dt = - \sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W/\partial \theta_x \quad (181)$$

or

$$m \sec \beta_{v xv x} \cos \beta_{tt}^x d/dt(\cos \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt) \quad (182)$$

$$= \sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W/\partial \theta_x$$

with  $\partial W/\partial \theta_x \geq 0$ . Combining equations (62) and (161) gives for attractive forces Newton's law as

$$m \csc \beta_{v xv x} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \quad (183)$$

$$= \sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W/\partial \theta_x$$

where the general expression for  $\beta_{v xv x}$  is given by equation (48) and  $d\theta_{vx}/dt$  is given by equation (56). The phase angle equation for this general case is given by equations (63) and (163) as

$$\theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) - \pi = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \quad (184)$$

which is valid for attractive forces.

The limiting case of a slow dynamical process in coherent space and incoherent time is obtained from equations (180), (182) and (183) which give for



attractive forces

$$m x \sec \beta_{vxvx}^{sci} d^2 \theta_x / dt^2 = \sec \beta_{ww} x^{-1} \partial W / \partial \theta_x \quad (185A)$$

$$m x \csc \beta_{vxvx}^{sci} (d\theta_x / dt)^2 = \sec \beta_{ww} x^{-1} \partial W / \partial \theta_x \quad (185B)$$

Combining equations (68) through (71) with equation (185) gives Newton's law for attractive forces as

$$m x [(E_{xt}^{sci})^2 + (F_{xt}^{sci})^2]^{1/2} = \sec \beta_{ww} x^{-1} \partial W / \partial \theta_x \quad (186)$$

where for attractive forces  $\partial W / \partial \theta_x \geq 0$ , and where  $E_{xt}^{sci}$  and  $F_{xt}^{sci}$  have the same forms as in equations (69) and (70) but with the sign of  $F_{xt}^{sci}$  reversed

$$E_{xt}^{sci} = (d\theta_x / dt)^2 \quad E_{xt}^{sci} \geq 0 \quad (187)$$

$$F_{xt}^{sci} = d^2 \theta_x / dt^2 \quad F_{xt}^{sci} \geq 0 \quad (188)$$

The case of a slow process in coherent space and incoherent time is obtained by combining equations (179) and (186) which gives for attractive forces the following form of Newton's law

$$m x [(E_{xt}^{sci})^2 + (F_{xt}^{sci})^2]^{1/2} = x^{-1} \partial W / \partial \theta_x \quad (189)$$

with  $\partial W / \partial \theta_x \geq 0$ . The phase angle equation for Newton's law of motion in coherent space and incoherent time is obtained from equations (63), (163) and (166) to be for slow attractive forces with  $\theta_w = 0$

$$\begin{aligned} \theta_{ax}^{ci+} &= \theta_x + \beta_{vxvx}^{sci} - \pi/2 \\ &= -\theta_x - \pi/2 \end{aligned} \quad (190)$$

where for this case  $\beta_{vxvx}^{sci}$  is given by equation (68)

$$\tan \beta_{vxvx}^{sci} = E_{xt}^{sci} / F_{xt}^{sci} \quad (191)$$

so that  $\beta_{vxvx}^{sci}$  is a small positive number. Equation (190) can also be written as

$$\beta_{vxvx}^{sci} + 2\theta_x = 0 \quad \theta_w = 0 \quad (192)$$

where for this case  $\beta_{vxvx}^{sci} \geq 0$  and  $\theta_x \leq 0$  and

$$\beta_{ww} \neq \beta_{xx} \quad (193)$$

The equivalent complex number form of Newton's dynamical law for a slow mechanical process in coherent space and incoherent time is

$$-m\bar{x}(E_{xt}^{sci} - jF_{xt}^{sci}) = -\partial\bar{W}/\partial\bar{x} = j/\bar{x} \partial W/\partial\theta_x \quad (194)$$

For repulsive forces equations (53) and (161) give Newton's law of motion as

$$m \sec \beta_{vxvx}^+ \cos \beta_{tt}^x \frac{dv_x}{dt} = -\sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W/\partial\theta_x \quad (195)$$

or equivalently

$$\begin{aligned} m \sec \beta_{vxvx}^+ \cos \beta_{tt}^x \frac{d}{dt}(\cos \beta_{tt}^x \csc \beta_{xx} x \frac{d\theta_x}{dt}) \\ = -\sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W/\partial\theta_x \end{aligned} \quad (196)$$

Combining equations (55) and (161) gives Newton's law of motion for repulsive forces as

$$\begin{aligned} m \csc \beta_{vxvx}^+ \cos^2 \beta_{tt}^x \csc \beta_{xx} x \frac{d\theta_x}{dt} \frac{d\theta_{vx}}{dt} \\ = -\sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W/\partial\theta_x \end{aligned} \quad (197)$$

with  $\partial W/\partial\theta_x \leq 0$ . For repulsive forces equations (57) and (163) give

$$\theta_x + \beta_{xx} + \beta_{vxvx}^+ - 2(\theta_t^x + \beta_{tt}^x) = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} + \pi \quad (198)$$

For the limiting case of a slow mechanical process that occurs in coherent space and incoherent time equations (196) and (197) for repulsive forces give the following form of Newton's law of motion

$$mx \sec \beta_{vxvx}^{sci+} d^2\theta_x/dt^2 = -x^{-1} \partial W/\partial\theta_x \quad (199A)$$

$$mx \csc \beta_{vxvx}^{sci+} (d\theta_x/dt)^2 = -x^{-1} \partial W/\partial\theta_x \quad (199B)$$

where

$$\tan \beta_{vxvx}^{sci+} = E_{xt}^{sci+}/F_{xt}^{sci+} \quad (200)$$

$$E_{xt}^{sci+} = E_{xt}^{sci} = (d\theta_x/dt)^2 \quad E_{xt}^{sci} \geq 0 \quad (201)$$

$$F_{xt}^{sci+} = F_{xt}^{sci} = d^2\theta_x/dt^2 \quad F_{xt}^{sci} \geq 0 \quad (202)$$

$$\beta_{vxvx}^{sci+} = \beta_{vxvx}^{sci} \quad (203)$$

and equation (200) gives Newton's law of motion for a repulsive force as

$$m\ddot{x}[(E_{xt}^{sci+})^2 + (F_{xt}^{sci+})^2]^{1/2} = -x^{-1}\partial W/\partial\theta_x \quad (204)$$

with  $\partial W/\partial\theta_x \leq 0$ , and equations (198) and (194) give Newton's law of motion for repulsive forces in coherent space and incoherent time as

$$2\theta_x + \beta_{vxvx}^{sci} = 0 \quad \theta_W = 0 \quad (205)$$

$$m\ddot{x}(E_{xt}^{sci+} - jF_{xt}^{sci+}) = -\partial\bar{W}/\partial\bar{x} = j/\bar{x} \partial W/\partial\theta_x \quad (206)$$

Equivalently Newton's law of motion in equation (206) for repulsive forces is written as

$$m\ddot{x}(E_{xt}^{sci} - jF_{xt}^{sci}) = -\partial\bar{W}/\partial\bar{x} = j/\bar{x} \partial W/\partial\theta_x \quad (207)$$

which is valid for repulsive forces acting in coherent space and incoherent time.

### Case 3. Slow Process, Incoherent Space and Coherent Time.

In this section a form of Newton's dynamical law is developed for broken symmetry spacetime that can be used to attain the case of a slow mechanical process in incoherent space and coherent time which is described by

$$\theta_W = 0 \quad \beta_{WW} = 0 \quad (208)$$

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \beta_{tt}^x = \pi/2 \quad (209)$$

where  $x$  and  $\theta_t^x$  are variables. Combining equations (95) and (159) gives Newton's law for an attractive force as

$$-m \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x = -\cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \quad (210)$$

From equations (96) and (159) it follows that equation (210) can be written as

$$\begin{aligned} m \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \\ = \cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \end{aligned} \quad (211)$$

Combining equations (90), (93) and (159) gives Newton's dynamical law for attractive forces as

$$-m \csc \beta_{vxvx} \sin \beta_{tt}^x v_x/t d\theta_{vx}/d\theta_t^x = -\cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \quad (212)$$

or equivalently

$$\begin{aligned} m \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} dx/d\theta_t^x d\theta_{vx}/d\theta_t^x \\ = \cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \end{aligned} \quad (213)$$

Equations (210) through (213) are valid for attractive forces for which  $\partial W/\partial x \geq 0$ . The phase angle condition for Newton's law of motion for this case is obtained from equations (97) and (163) as

$$\theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) + \pi = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \quad (214)$$

which is valid for attractive forces.

For the limiting case of a slowly changing potential in incoherent space and coherent time Newton's dynamical equations (211) and (213) for attractive forces become

$$m t^{-2} \sec \beta_{v xv x}^{sic} d^2 x / d\theta_t^{x2} = \partial W / \partial x \quad (215)$$

$$m t^{-2} \csc \beta_{v xv x}^{sic} (-dx / d\theta_t^x) = \partial W / \partial x \quad (216)$$

For this case Newton's dynamical laws (211) and (213) can also be written for attractive forces as

$$m t^{-2} [(E_{xt}^{sic})^2 + (F_{xt}^{sic})^2]^{1/2} = \partial W / \partial x \quad (217)$$

with  $\partial W / \partial x \geq 0$  for an attractive force, and where

$$\tan \beta_{v xv x}^{sic} = E_{xt}^{sic} / F_{xt}^{sic} \quad (218)$$

$$E_{xt}^{sic} = E_{xt}^{ic} = -dx / d\theta_t^x \quad E_{xt}^{sic} \geq 0 \quad (219)$$

$$F_{xt}^{sic} = F_{xt}^{ic} = d^2 x / d\theta_t^{x2} \quad F_{xt}^{sic} \geq 0 \quad (220)$$

where  $E_{xt}^{ic}$  and  $F_{xt}^{ic}$  are given by equations (103) and (104). It is clear that

$$\beta_{v xv x}^{sic} = \beta_{v xv x}^{ic} \quad (221)$$

where  $\beta_{v xv x}^{ic}$  is a small number given by equation (102). The corresponding phase angle condition obtained from equation (214) is

$$\beta_{v xv x}^{sic} - 2\theta_t^x = 0 \quad \theta_W = 0 \quad (222)$$

The equivalent complex number form of Newton's law of motion is written for attractive forces as

$$-m/\bar{t}^2 (F_{xt}^{sic} + jE_{xt}^{sic}) = -\partial \bar{W} / \partial \bar{x} = -\partial W / \partial x \quad (223)$$

which holds for slow attractive forces in incoherent space and coherent time.

In the case of repulsive forces Newton's dynamical equations (87) and (159) give

$$m \sec \beta_{vxvx}^+ \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x = - \cos \beta_{xx} \sec \beta_{ww} \partial W/\partial x \quad (224)$$

while equations (89) and (159) give Newton's law for repulsive forces as

$$\begin{aligned} m \sec \beta_{vxvx}^+ \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \\ = - \cos \beta_{xx} \sec \beta_{ww} \partial W/\partial x \end{aligned} \quad (225)$$

Combining equations (91) and (159) gives Newton's dynamical law for repulsive forces as

$$\begin{aligned} m \csc \beta_{vxvx}^+ \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} dx/d\theta_t^x d\theta_{vx}/d\theta_t^x \\ = - \cos \beta_{xx} \sec \beta_{ww} \partial W/\partial x \end{aligned} \quad (226)$$

where for a repulsive force  $\partial W/\partial x \leq 0$ . For a repulsive force equations (88) and (163) give

$$\theta_x + \beta_{xx} + \beta_{vxvx}^+ - 2(\theta_t^x + \beta_{tt}^x) = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} - \pi \quad (227)$$

In the limiting case of a slow mechanical process in incoherent space and coherent time Newton's dynamical equations (225) and (226) become for repulsive forces

$$mt^{-2} \sec \beta_{vxvx}^{sic+} d^2 x/d\theta_t^{x2} = - \partial W/\partial x \quad (228)$$

$$mt^{-2} \csc \beta_{vxvx}^{sic+} (- dx/d\theta_t^x) = - \partial W/\partial x \quad (229)$$

where

$$\tan \beta_{vxvx}^{sic+} = E_{xt}^{sic+}/F_{xt}^{sic+} \quad (230)$$

$$E_{xt}^{sic+} = E_{xt}^{sic} = E_{xt}^{ic} \quad E_{xt}^{sic+} \geq 0 \quad (231)$$

$$F_{xt}^{sic+} = F_{xt}^{sic} = F_{xt}^{ic} \quad F_{xt}^{sic+} \geq 0 \quad (232)$$

$$\beta_{vxvx}^{sic+} = \beta_{vxvx}^{sic} \quad (233)$$

and Newton's dynamical law for repulsive forces in equations (228) and (229) becomes

$$m t^{-2} [(E_{xt}^{sic+})^2 + (F_{xt}^{sic+})^2]^{1/2} = - \partial W / \partial x \quad (234)$$

Equivalently, equation (234) can be written as

$$m t^{-2} [(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} = - \partial W / \partial x \quad (235)$$

with  $\partial W / \partial x \leq 0$  for a repulsive force. The phase angle equation (227) for Newton's dynamical law with repulsive forces becomes with  $\theta_W = 0$

$$\beta_{vxvx}^{sic} - 2\theta_t^x = 0 \quad \theta_t^x \geq 0 \quad (236)$$

or

$$\beta_{vxvx}^{ic} - 2\theta_t^x = 0 \quad \theta_t^x \geq 0 \quad (237)$$

Finally, the complex number form of Newton's law of motion for a slow mechanical process in incoherent space and coherent time and for a repulsive force is written in an analogous fashion to equation (223) as

$$m/t^{-2} (F_{xt}^{sic+} + jE_{xt}^{sic+}) = - \partial \bar{W} / \partial \bar{x} = - \partial W / \partial x \quad (238)$$

Equation (238) can be written in the following two equivalent forms of Newton's dynamical law of motion for this special case

$$m/t^{-2} (F_{xt}^{sic} + jE_{xt}^{sic}) = - \partial W / \partial x \quad (239)$$

$$m/t^{-2} (F_{xt}^{ic} + jE_{xt}^{ic}) = - \partial W / \partial x \quad (240)$$

where  $\theta_W = 0$ , and where for a repulsive force  $\partial W / \partial x \leq 0$ .

Case 4. Slow Process, Coherent Space and Coherent Time.

A form of Newton's law is considered that can be reduced to the special case of a slow mechanical process in coherent space and coherent time which is described by

$$\beta_{WW} = 0 \quad \theta_W = 0 \quad (241)$$

$$\beta_{xx} = \pi/2 \quad \beta_{tt}^x = \pi/2 \quad (242)$$

where  $\theta_x$  and  $\theta_t^x$  are variables. A form of Newton's law that can be used for this case comes from equations (36), (124) and (161) with the result that for an attractive force

$$- m \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} dv_x / d\theta_t^x = - \sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W / \partial \theta_x \quad (243)$$

or combining equations (21) and (243) gives Newton's dynamical law for an attractive force as

$$\begin{aligned}
& m \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \\
& = \sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W/\partial \theta_x
\end{aligned} \tag{244}$$

where for an attractive force  $\partial W/\partial \theta_x \geq 0$ . Another form of Newton's law that is suitable for this case is obtained from equations (118), (123), (124) and (161) which give for an attractive force

$$\begin{aligned}
& - m \csc \beta_{vxvx} \sin \beta_{tt}^x v_x/t d\theta_{vx}/d\theta_t^x \\
& = - \sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W/\partial \theta_x
\end{aligned} \tag{245}$$

Equation (245) can be rewritten using equation (126) to give the following form of Newton's dynamical law for attractive forces

$$\begin{aligned}
& mxt^{-2} \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} d\theta_x/d\theta_t^x d\theta_{vx}/d\theta_t^x \\
& = \sin \beta_{xx} \sec \beta_{ww} x^{-1} \partial W/\partial \theta_x
\end{aligned} \tag{246}$$

where  $d\theta_{vx}/d\theta_t^x$  is given by equation (121). For an attractive force  $\partial W/\partial \theta_x \geq 0$ . The corresponding phase angle equations (128), (163) and (168) give the phase angle relation

$$\theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) + \pi = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} \tag{247}$$

which is valid for a slow mechanical process in nearly coherent space and nearly coherent time.

The case of a slow mechanical process that occurs in coherent space and coherent time follows from equations (121), (241), (242), (244) and (246) which give Newton's dynamical law as

$$mxt^{-2} \sec \beta_{vxvx}^{scc} d^2 \theta_x/d\theta_t^{x2} = x^{-1} \partial W/\partial \theta_x \tag{248}$$

$$mxt^{-2} \csc \beta_{vxvx}^{scc} d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) = x^{-1} \partial W/\partial \theta_x \tag{249}$$

Equations (248) and (249) can be written as

$$mxt^{-2} [(E_{xt}^{scc})^2 + (F_{xt}^{scc})^2]^{1/2} = x^{-1} \partial W/\partial \theta_x \tag{250}$$

where for an attractive force  $\partial W/\partial \theta_x \geq 0$ , and where

$$\tan \beta_{vxvx}^{scc} = E_{xt}^{scc}/F_{xt}^{scc} \tag{251}$$

$$E_{xt}^{scc} = d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad E_{xt}^{scc} \leq 0 \quad (252)$$

$$F_{xt}^{scc} = d^2\theta_x/d\theta_t^{x2} \quad F_{xt}^{scc} \leq 0 \quad (253)$$

It follows that

$$\beta_{vxvx}^{scc} = -\pi + \alpha_{xt} \quad (254)$$

where the small positive number  $\alpha_{xt}$  is given by

$$\tan \alpha_{xt} = |E_{xt}^{scc}|/|F_{xt}^{scc}| \quad (255)$$

For this case the phase angle equation (247) becomes with  $\theta_W = 0$

$$\theta_x - 2\theta_t^x + \beta_{vxvx}^{scc} + \pi/2 = -\theta_x - \pi/2 \quad (256)$$

Equation (256) can be rewritten with the help of equation (254) as

$$2(\theta_x - \theta_t^x) + \alpha_{xt} = 0 \quad \theta_t^x \geq \theta_x \quad (257)$$

The complex number form of Newton's dynamical law that is equivalent to equations (248) through (257) is written as for attractive forces as

$$-m\bar{x}/t^2(-E_{xt}^{scc} + jF_{xt}^{scc}) = -\partial\bar{W}/\partial\bar{x} = j/\bar{x} \partial W/\partial\theta_x \quad (258)$$

where  $\theta_W = 0$  for a slow mechanical process, and  $\partial W/\partial\theta_x \geq 0$  for an attractive force.

For a repulsive force equations (122) and (161) give the form of Newton's dynamical law as

$$\begin{aligned} m \sec \beta_{vxvx}^+ \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \\ = -\sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W/\partial\theta_x \end{aligned} \quad (259)$$

while equations (120) and (161) give for a repulsive force

$$\begin{aligned} mxt^{-2} \csc \beta_{vxvx}^+ \sin^2 \beta_{tt}^x \csc \beta_{xx} d\theta_x/d\theta_t^x d\theta_{vx}/d\theta_t^x \\ = -\sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W/\partial\theta_x \end{aligned} \quad (260)$$

with  $\partial W/\partial\theta_x \leq 0$  for repulsive forces. For a repulsive force equations (119) and (163) give

$$\theta_x + \beta_{xx} + \beta_{vxvx}^+ - 2(\theta_t^x + \beta_{tt}^x) = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} - \pi \quad (261)$$



In the limiting case of a slowly varying repulsive potential in coherent space and coherent time equations (259) and (260) become

$$mxt^{-2} \sec \beta_{vxvx}^{scc+} d^2\theta_x/d\theta_t^{x2} = -x^{-1} \partial W/\partial \theta_x \quad (262A)$$

$$mxt^{-2} \csc \beta_{vxvx}^{scc+} d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) = -x^{-1} \partial W/\partial \theta_x \quad (262B)$$

where

$$\tan \beta_{vxvx}^{scc+} = E_{xt}^{scc+}/F_{xt}^{scc+} \quad (263)$$

$$E_{xt}^{scc+} = E_{xt}^{scc} \quad E_{xt}^{scc+} \leq 0 \quad (264)$$

$$F_{xt}^{scc+} = F_{xt}^{scc} \quad F_{xt}^{scc+} \leq 0 \quad (265)$$

$$\beta_{vxvx}^{scc+} = \beta_{vxvx}^{scc} = -\pi + \alpha_{xt} \quad (266)$$

Equation (262) for repulsive forces can be rewritten as

$$mxt^{-2} [(E_{xt}^{scc})^2 + (F_{xt}^{scc})^2]^{1/2} = -x^{-1} \partial W/\partial \theta_x \quad (267)$$

with  $\partial W/\partial \theta_x \leq 0$ , while equation (261) becomes with  $\theta_W = 0$

$$\theta_x - 2\theta_t^x + \alpha_{xt} = -\theta_x \quad (268)$$

or equivalently

$$2(\theta_x - \theta_t^x) + \alpha_{xt} = 0 \quad \theta_t^x \geq \theta_x \quad (269)$$

For a repulsive force the complex number form of Newton's dynamical law that is equivalent to equations (259) through (269) is written as

$$m\bar{x}/\bar{t}^2 (-E_{xt}^{scc} + jF_{xt}^{scc}) = -\partial \bar{W}/\partial \bar{x} = j/\bar{x} \partial W/\partial \theta_x \quad (270)$$

where  $\partial W/\partial \theta_x \leq 0$  for a repulsive force.

#### Case 5. Ultrafast Process, Incoherent Space and Incoherent Time.

Consider now the form of Newton's law of motion that can be used to regain the case of an ultrafast process in incoherent space and incoherent time that is described by

$$\beta_{WW} = \pi/2 \quad (271)$$

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (272)$$

with  $x$  and  $t$  as variables. Combining equations (40), (46), (47), (51), (156), (160), (163) and (165) gives Newton's dynamical law as

$$m \cos \beta_{tt}^x \sec \beta_{v xv x} dv_x/dt = - \cos \beta_{xx} \csc \beta_{WW} W \partial \theta_W / \partial x \quad (273)$$

$$\begin{aligned} \theta_{ax} &= \theta_{vx} + \beta_{v xv x} - \theta_t^x - \beta_{tt}^x \\ &= \theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) \\ &= \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \end{aligned} \quad (274)$$

with  $\beta_{v xv x}$  and  $\beta_{WW}$  given by equations (48) and (164) respectively. Combining equations (18) and (273) gives Newton's dynamical law as

$$\begin{aligned} m \cos \beta_{tt}^x \sec \beta_{v xv x} d/dt(\cos \beta_{tt}^x \sec \beta_{xx} dx/dt) \\ = - \cos \beta_{xx} \csc \beta_{WW} W \partial \theta_W / \partial x \end{aligned} \quad (275)$$

Combining equations (50) and (160) gives an alternative form of Newton's law of motion as

$$m \csc \beta_{v xv x} \cos \beta_{tt}^x v_x d\theta_{vx}/dt = - \cos \beta_{xx} \csc \beta_{WW} W \partial \theta_W / \partial x \quad (276)$$

or equivalently

$$\begin{aligned} m \csc \beta_{v xv x} \cos^2 \beta_{tt}^x \sec \beta_{xx} dx/dt d\theta_{vx}/dt \\ = - \cos \beta_{xx} \csc \beta_{WW} W \partial \theta_W / \partial x \end{aligned} \quad (277)$$

with  $W \partial \theta_W / \partial x \geq 0$  for an attractive force, and  $W \partial \theta_W / \partial x \leq 0$  for a repulsive force.

The limiting case of motion in incoherent space and incoherent time is obtained by combining equations (272) and (275) with the result that Newton's dynamical law becomes

$$m d^2 x/dt^2 = - \csc \beta_{WW} W \partial \theta_W / \partial x \quad (278)$$

The further limiting case of an ultrafast dynamical system in incoherent space and incoherent time is derived by combining equations (271) and (278) which gives Newton's law of motion as

$$m d^2 x/dt^2 = - W \partial \theta_W / \partial x \quad (279)$$

The phase angle equation for Newton's law of motion in this case is obtained by combining equations (271), (272) and (274) with the result

$$\theta_{ax}^{uii} = \beta_{vxvx}^{uii} = \theta_W + \pi/2 \quad (280)$$

where equations (279) and (280) are valid for attractive and repulsive forces. For an attractive force  $W\partial\theta_W/\partial x \geq 0$  while for a repulsive force  $W\partial\theta_W/\partial x \leq 0$ .

#### Case 6. Ultrafast Process, Coherent Space and Incoherent Time.

This section develops a form of Newton's dynamical law that is suitable for making the transition to the case of an ultrafast dynamical system in coherent space and incoherent time which is described by

$$\beta_{WW} = \pi/2 \quad (281)$$

$$\beta_{xx} = \pi/2 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (282)$$

Combining equations (53), (59) and (162) gives for an attractive force the following form of Newton's law of motion

$$-m \sec \beta_{vxvx} \cos \beta_{tt}^x dv_x/dt = -\sin \beta_{xx} \csc \beta_{WW} W/x \partial\theta_W/\partial\theta_x \quad (283)$$

Combining equations (19) and (283) gives Newton's law as

$$\begin{aligned} m \sec \beta_{vxvx} \cos \beta_{tt}^x d/dt(\cos \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt) \\ = \sin \beta_{xx} \csc \beta_{WW} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (284)$$

where for an attractive force  $W\partial\theta_W/\partial\theta_x \geq 0$ . The combination of equations (61) and (162) gives an alternative form of Newton's law of motion for an attractive force as follows

$$-m \csc \beta_{vxvx} \cos \beta_{tt}^x v_x d\theta_{vx}/dt = -\sin \beta_{xx} \csc \beta_{WW} W/x \partial\theta_W/\partial\theta_x \quad (285)$$

Combining equations (62) and (162) or combining equations (19) and (285) gives Newton's dynamical law for attractive forces as

$$\begin{aligned} m \csc \beta_{vxvx} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \\ = \sin \beta_{xx} \csc \beta_{WW} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (286)$$

where  $d\theta_{vx}/dt$  is given by equation (56). The phase angle equation for the acceleration can be obtained from equations (63) and (163) as

$$\theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) - \pi = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \quad (287)$$

which is valid for a fast process in nearly coherent space and nearly incoherent time.

For the limiting case of an attractive ultrafast mechanical force in coherent space and incoherent time, equations (281), (282), (284) and (286) give Newton's dynamical law as

$$m_x \sec \beta_{vxvx}^{uci} d^2\theta_x/dt^2 = W/x \partial\theta_W/\partial\theta_x \quad (288)$$

$$m_x \csc \beta_{vxvx}^{uci} (d\theta_x/dt)^2 = W/x \partial\theta_W/\partial\theta_x \quad (289)$$

Equations (288) and (289) can be rewritten as the following form of Newton's law of motion

$$m_x [(E_{xt}^{uci})^2 + (F_{xt}^{uci})^2]^{1/2} = W/x \partial\theta_W/\partial\theta_x \quad (290)$$

where for an attractive force  $W\partial\theta_W/\partial\theta_x \geq 0$ , and where

$$\tan \beta_{vxvx}^{uci} = E_{xt}^{uci}/F_{xt}^{uci} \quad (291)$$

$$E_{xt}^{uci} = E_{xt}^{ci} = (d\theta_x/dt)^2 \quad E_{xt}^{uci} \geq 0 \quad (292)$$

$$F_{xt}^{uci} = F_{xt}^{ci} = d^2\theta_x/dt^2 \quad F_{xt}^{uci} \leq 0 \quad (293)$$

$$\beta_{vxvx}^{uci} = \beta_{vxvx}^{ci} = \pi/2 + \kappa_{xt} \quad (294)$$

where  $\beta_{vxvx}^{ci}$  is given by equation (68) and  $\kappa_{xt}$  is given by equation (73). The internal phase angle equation for Newton's law of motion for this case follows from equations (281), (282), (287) and (294) as

$$\theta_x + \beta_{vxvx}^{uci} - \pi/2 = \theta_W - \theta_x \quad (295)$$

or equivalently

$$\theta_W = 2\theta_x + \kappa_{xt} \quad (296)$$

The equivalent complex number form of Newton's law of motion for an attractive force in this special case is given by

$$-m\bar{x}(E_{xt}^{uci} - jF_{xt}^{uci}) = -\partial\bar{W}/\partial\bar{x} = -\bar{W}/\bar{x} \partial\theta_W/\partial\theta_x \quad (297)$$

with  $W\partial\theta_W/\partial\theta_x \geq 0$  for an attractive force.

For a repulsive force equations (53) and (162) give Newton's law of motion as

$$\begin{aligned}
& m \sec \beta_{vxvx}^+ \cos \beta_{tt}^x \frac{d}{dt} (\cos \beta_{tt}^x \csc \beta_{xx} \times d\theta_x / dt) \\
& = - \sin \beta_{xx} \csc \beta_{ww} W/x \partial \theta_w / \partial \theta_x
\end{aligned} \tag{298}$$

Alternatively, combining equations (55) and (162) gives Newton's dynamical law for repulsive forces as

$$\begin{aligned}
& m \csc \beta_{vxvx}^+ \cos^2 \beta_{tt}^x \csc \beta_{xx} \times d\theta_x / dt \, d\theta_{vx} / dt \\
& = - \sin \beta_{xx} \csc \beta_{ww} W/x \partial \theta_w / \partial \theta_x
\end{aligned} \tag{299}$$

with  $W \partial \theta_w / \partial \theta_x \leq 0$ . For repulsive forces equations (57) and (163) give the phase angle condition as

$$\theta_x + \beta_{xx} + \beta_{vxvx}^+ - 2(\theta_t^x + \beta_{tt}^x) = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} + \pi \tag{300}$$

In the case of an ultrafast mechanical process that occurs in coherent space and incoherent time, equations (298) and (299) become for repulsive forces the following Newtonian laws of motion

$$m x \sec \beta_{vxvx}^{uci+} d^2 \theta_x / dt^2 = - W/x \partial \theta_w / \partial \theta_x \tag{301}$$

$$m x \csc \beta_{vxvx}^{uci+} (d\theta_x / dt)^2 = - W/x \partial \theta_w / \partial \theta_x \tag{302}$$

where  $W \partial \theta_w / \partial \theta_x \leq 0$  for a repulsive force, and where

$$\tan \beta_{vxvx}^{uci+} = E_{xt}^{uci+} / F_{xt}^{uci+} \tag{303}$$

with

$$E_{xt}^{uci+} = E_{xt}^{uci} = (d\theta_x / dt)^2 \quad E_{xt}^{uci+} \geq 0 \tag{304}$$

$$F_{xt}^{uci+} = F_{xt}^{uci} = d^2 \theta_x / dt^2 \quad F_{xt}^{uci+} \leq 0 \tag{305}$$

Then it follows that

$$\beta_{vxvx}^{uci+} = \beta_{vxvx}^{uci} = \pi/2 + \kappa_{xt} \tag{306}$$

where  $\kappa_{xt}$  is given by equation (73). Then equations (301) and (302) can be written as

$$m x [(E_{xt}^{uci+})^2 + (F_{xt}^{uci+})^2]^{1/2} = - W/x \partial \theta_w / \partial \theta_x \tag{307}$$

which is valid for repulsive forces with  $W \partial \theta_w / \partial \theta_x \leq 0$ . Equation (300) becomes

$$\theta_x + \kappa_{xt} = \theta_w - \theta_x \tag{308}$$

Then the equivalent of equation (297) for repulsive forces is

$$m\bar{x}(E_{xt}^{uc1+} - jF_{xt}^{uc1+}) = -\partial\bar{W}/\partial\bar{x} = -\bar{W}/\bar{x} \partial\theta_W/\partial\theta_x \quad (309)$$

with  $W\partial\theta_W/\partial\theta_x \leq 0$ .

#### Case 7. Ultrafast Process, Incoherent Space and Coherent Time.

A formulation of Newton's dynamical law is presented that correctly reduces to the limiting case of an ultrafast mechanical process in incoherent space and coherent time which is described by

$$\beta_{WW} = \pi/2 \quad (310)$$

$$\beta_{xx} = 0 \quad \beta_{tt}^x = \pi/2 \quad (311)$$

where  $x$  and  $\theta_t^x$  are variables. Combining equations (95) and (160) gives Newton's law of motion for attractive forces as

$$-m \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x = -\cos \beta_{xx} \csc \beta_{WW} W \partial\theta_W/\partial x \quad (312)$$

Using equations (20) and (96) allows equation (312) to be written as

$$\begin{aligned} m \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \\ = \cos \beta_{xx} \csc \beta_{WW} W \partial\theta_W/\partial x \end{aligned} \quad (313)$$

with  $W\partial\theta_W/\partial x \geq 0$ . An alternative form of Newton's law of motion for this situation comes from equations (90), (91) and (160) with the result that for an attractive force

$$-m \csc \beta_{vxvx} \sin \beta_{tt}^x v_x/t d\theta_{vx}/d\theta_t^x = -\cos \beta_{xx} \csc \beta_{WW} W \partial\theta_W/\partial x \quad (314)$$

or equivalently using equations (20) and (314) gives Newton's dynamical law as

$$\begin{aligned} m \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} dx/d\theta_t^x d\theta_{vx}/d\theta_t^x \\ = \cos \beta_{xx} \csc \beta_{WW} W \partial\theta_W/\partial x \end{aligned} \quad (315)$$

with  $W\partial\theta_W/\partial x \geq 0$  for an attractive force. Equations (312) through (315) are completely general. The phase angle equation for this general case can be written using equations (97) and (163) as

$$\theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) + \pi = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \quad (316)$$

which can be used to attain the limiting case of interest.

The limiting case of an ultrafast attractive potential acting in incoherent space and coherent time is obtained from equations (310), (311), (313) and (315) as the following forms of Newton's dynamical law

$$m t^{-2} \sec \beta_{vxvx}^{uic} d^2 x / d\theta_t^2 = W \partial \theta_W / \partial x \quad (317)$$

$$m t^{-2} \csc \beta_{vxvx}^{uic} (-dx / d\theta_t^x) = W \partial \theta_W / \partial x \quad (318)$$

Equations (317) and (318) can also be written for an attractive force to give Newton's law of motion as

$$m t^{-2} [(E_{xt}^{uic})^2 + (F_{xt}^{uic})^2]^{1/2} = W \partial \theta_W / \partial x \quad (319)$$

with  $W \partial \theta_W / \partial x \geq 0$ , and where

$$\tan \beta_{vxvx}^{uic} = E_{xt}^{uic} / F_{xt}^{uic} \quad (320)$$

$$E_{xt}^{uic} = -dx / d\theta_t^x \quad E_{xt}^{uic} \geq 0 \quad (321)$$

$$F_{xt}^{uic} = d^2 x / d\theta_t^2 \quad F_{xt}^{uic} \geq 0 \quad (322)$$

$$\beta_{vxvx}^{uic} = \pi/2 - \gamma_{xt} \quad (323)$$

where

$$\tan \gamma_{xt} = F_{xt}^{uic} / E_{xt}^{uic} \quad (324)$$

and  $\gamma_{xt}$  is a small positive number. For this case the phase angle condition (316) is written as

$$\beta_{vxvx}^{uic} - 2\theta_t^x = \theta_W + \pi/2 \quad (325)$$

or equivalently as

$$\theta_W = -\gamma_{xt} - 2\theta_t^x \quad (326)$$

The complex number form of Newton's law of motion for an attractive ultrafast potential in incoherent space and coherent time is given by

$$-m/t^2 (F_{xt}^{uic} + j E_{xt}^{uic}) = -\partial \bar{W} / \partial \bar{x} = -j \bar{W} \partial \theta_W / \partial x \quad (327)$$

with  $W \partial \theta_W / \partial x \geq 0$ .

In the presence of a repulsive force, equations (89) and (160) give Newton's equations of motion as

$$m \sec \beta_{vxvx}^+ \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \quad (328)$$

$$= - \cos \beta_{xx} \csc \beta_{ww} W \partial \theta_w / \partial x$$

Combining equations (91) and (160) for repulsive forces gives Newton's law as

$$m \csc \beta_{vxvx}^+ \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} dx/d\theta_t^x d\theta_{vx}/d\theta_t^x \quad (329)$$

$$= - \cos \beta_{xx} \csc \beta_{ww} W \partial \theta_w / \partial x$$

with  $W \partial \theta_w / \partial x \leq 0$  for a repulsive force. For a repulsive force equations (88) and (163) give the phase angle condition as

$$\theta_x + \beta_{xx} + \beta_{vxvx}^+ - 2(\theta_t^x + \beta_{tt}^x) = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} - \pi \quad (330)$$

For the case of an ultrafast mechanical process in incoherent space and coherent time equations (311), (328) and (329) become

$$mt^{-2} \sec \beta_{vxvx}^{uic+} d^2 x / d\theta_t^{x2} = - W \partial \theta_w / \partial x \quad (331)$$

$$mt^{-2} \csc \beta_{vxvx}^{uic+} (- dx / d\theta_t^x) = - W \partial \theta_w / \partial x \quad (332)$$

where  $W \partial \theta_w / \partial x \leq 0$  for repulsive forces, and where

$$\tan \beta_{vxvx}^{uic+} = E_{xt}^{uic+} / F_{xt}^{uic+} \quad (333)$$

$$E_{xt}^{uic+} = E_{xt}^{uic} = - dx / d\theta_t^x \quad E_{xt}^{uic+} \geq 0 \quad (334)$$

$$F_{xt}^{uic+} = F_{xt}^{uic} = d^2 x / d\theta_t^{x2} \quad F_{xt}^{uic+} \geq 0 \quad (335)$$

$$\beta_{vxvx}^{uic+} = \beta_{vxvx}^{uic} = \pi/2 - \gamma_{xt} \quad (336)$$

where  $\gamma_{xt}$  is given by equation (324). Equations (331) and (332) can be written for repulsive forces as

$$mt^{-2} [(E_{xt}^{uic+})^2 + (F_{xt}^{uic+})^2]^{1/2} = - W \partial \theta_w / \partial x \quad (337)$$

while the phase angle equation (330) becomes

$$\theta_w = - \gamma_{xt} - 2\theta_t^x \quad (338)$$

The corresponding complex number form of Newton's law of motion for a repulsive ultrafast potential acting in incoherent space and coherent time is



$$m/t^2(F_{xt}^{uic} + jE_{xt}^{uic}) = -\partial\bar{W}/\partial\bar{x} = -j\bar{W}\partial\theta_W/\partial x \quad (339)$$

with  $W\partial\theta_W/\partial x \leq 0$  and  $\theta_x = 0$  for a repulsive force in incoherent space.

#### Case 8. Ultrafast Process, Coherent Space and Coherent Time.

This section considers a formulation of Newton's law of dynamics in broken symmetry space and time that reduces to the limiting case of an ultrafast mechanical process in totally coherent spacetime which is described by

$$\beta_{WW} = \pi/2 \quad (340)$$

$$\beta_{xx} = \pi/2 \quad \beta_{tt}^x = \pi/2 \quad (341)$$

where  $\theta_x$  and  $\theta_t^x$  are variables. Equations (36), (124) and (162) give Newton's law of motion for an attractive force as

$$-m \sec \beta_{vxx} \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x = -\csc \beta_{WW} \sin \beta_{xx} W/x \partial\theta_W/\partial\theta_x \quad (342)$$

or equivalently combining equations (21) and (342) gives Newton's law for an attractive force as

$$\begin{aligned} & m t^{-1} \sec \beta_{vxx} \sin \beta_{tt}^x d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \\ & = \csc \beta_{WW} \sin \beta_{xx} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (343)$$

Another form of Newton's law of motion for this case is obtained from equations (118), (123), (124), (156) and (162) which gives the following result for an attractive force

$$\begin{aligned} & -m \csc \beta_{vxx} \sin \beta_{tt}^x v_x/t d\theta_{vx}/d\theta_t^x \\ & = -\csc \beta_{WW} \sin \beta_{xx} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (344)$$

Equation (344) can be rewritten using equations (21) or (126) as follows

$$\begin{aligned} & m x t^{-2} \csc \beta_{vxx} \sin^2 \beta_{tt}^x \csc \beta_{xx} d\theta_x/d\theta_t^x d\theta_{vx}/d\theta_t^x \\ & = \csc \beta_{WW} \sin \beta_{xx} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (345)$$

where  $W\partial\theta_W/\partial\theta_x \geq 0$  for the case of an attractive force. The corresponding phase angle equations (128), (163) and (168) give

$$\theta_x + \beta_{vxx} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) + \pi = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \quad (346)$$

which is valid for nearly coherent space and nearly coherent time.

The case of an ultrafast attractive mechanical potential in coherent space and coherent time is obtained from equations (340) and (341), and Newton's dynamical law for this special case is determined from equations (343) and (345) to be

$$mxt^{-2} \sec \beta_{vxvx}^{ucc} d^2\theta_x/d\theta_t^2 = W/x \partial\theta_W/\partial\theta_x \quad (347A)$$

$$mxt^{-2} \csc \beta_{vxvx}^{ucc} d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) = W/x \partial\theta_W/\partial\theta_x \quad (347B)$$

Newton's law of motion in equation (347) can also be written for this case as

$$mxt^{-2} [(E_{xt}^{ucc})^2 + (F_{xt}^{ucc})^2]^{1/2} = W/x \partial\theta_W/\partial\theta_x \quad (348)$$

where  $W\partial\theta_W/\partial\theta_x \geq 0$  for an attractive force, and where

$$\tan \beta_{vxvx}^{ucc} = E_{xt}^{ucc}/F_{xt}^{ucc} \quad (349)$$

$$E_{xt}^{ucc} = E_{xt}^c = d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad E_{xt}^c \leq 0 \quad (350)$$

$$F_{xt}^{ucc} = F_{xt}^c = d^2\theta_x/d\theta_t^2 \quad F_{xt}^c \geq 0 \quad (351)$$

where  $E_{xt}^c$  and  $F_{xt}^c$  are given by equations (135) and (136). Therefore

$$\tan \beta_{vxvx}^{ucc} = E_{xt}^{ucc}/F_{xt}^{ucc} \quad \beta_{vxvx}^{ucc} = \beta_{vxvx}^c = -\pi/2 + \delta_{xt} \quad (352)$$

where  $\beta_{vxvx}^c$  is given by equation (140). The phase angle condition for Newton's law of motion for an ultrafast mechanical process in coherent spacetime is obtained from equations (143), (163) and (168) or directly from equation (346) to be

$$\begin{aligned} \theta_{ax}^{c'} &= \theta_x + \beta_{vxvx}^{ucc} + \pi/2 - 2\theta_t^x \\ &= \theta_x - 2\theta_t^x + \delta_{xt} \\ &= \theta_W - \theta_x \end{aligned} \quad (353)$$

where  $\delta_{xt}$  is given by equations (140) and (141). Equivalently, equation (353) can be rewritten as

$$2(\theta_x - \theta_t^x) + \delta_{xt} = \theta_W \quad (354)$$

Equations (348) through (354) are equivalent to the complex number form of Newton's dynamical law of motion given by equation (156) which for an ultrafast attractive potential in coherent space and coherent time is written as

$$-m\bar{x}/\bar{t}^2 (-E_{xt}^{ucc} + jF_{xt}^{ucc}) = -\bar{W}/\bar{x} \partial\theta_W/\partial\theta_x \quad (355)$$

with  $W\partial\theta_W/\partial\theta_x > 0$ . For a free particle moving in coherent spacetime

$$\partial\theta_W/\partial\theta_x = 0 \quad (356)$$

and two possible solutions to equation (355) can be found

$$\theta_x = c_1 \quad (357)$$

$$\theta_x = \theta_t^x + c_2 \quad (358)$$

where  $c_1$  and  $c_2$  are constants. These solutions can also be deduced from equation (348). Equation (357) represents a state of rest for internal motion, and equation (358) represents a state of uniform motion in internal spacetime.

For a repulsive force, equations (122) and (236) give Newton's law of motion as

$$\begin{aligned} & m t^{-1} \sec \beta_{vxvx}^+ \sin \beta_{tt}^x \frac{d}{d\theta_t^x} (\sin \beta_{tt}^x \csc \beta_{xx} x/t \frac{d\theta_x}{d\theta_t^x}) \\ & = - \csc \beta_{WW} \sin \beta_{xx} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (359)$$

An alternative expression for Newton's law of motion for repulsive forces is obtained from equations (120) and (236) as

$$\begin{aligned} & m x t^{-2} \csc \beta_{vxvx}^+ \sin^2 \beta_{tt}^x \csc \beta_{xx} \frac{d\theta_x}{d\theta_t^x} \frac{d\theta_{vx}}{d\theta_t^x} \\ & = - \csc \beta_{WW} \sin \beta_{xx} W/x \partial\theta_W/\partial\theta_x \end{aligned} \quad (360)$$

where  $W\partial\theta_W/\partial\theta_x \leq 0$  for a repulsive force. For repulsive forces the phase angle equation for Newton's law of motion is obtained from equations (119) and (163) to be

$$\theta_x + \beta_{xx} + \beta_{vxvx}^+ - 2(\theta_t^x + \beta_{tt}^x) = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} - \pi \quad (361)$$

In the limiting case of an ultrafast repulsive mechanical potential in coherent space and coherent time equations (359) and (360) become

$$m x t^{-2} \sec \beta_{vxvx}^{ucc+} \frac{d^2 \theta_x}{d\theta_t^{x2}} = - W/x \partial\theta_W/\partial\theta_x \quad (362)$$

$$m x t^{-2} \csc \beta_{vxvx}^{ucc+} \frac{d\theta_x}{d\theta_t^x} \left( \frac{d\theta_x}{d\theta_t^x} - 1 \right) = - W/x \partial\theta_W/\partial\theta_x \quad (363)$$

where

$$\tan \beta_{vxvx}^{ucc+} = E_{xt}^{ucc+} / F_{xt}^{ucc+} \quad (364)$$

$$E_{xt}^{ucc+} = E_{xt}^{ucc} = E_{xt}^c \quad E_{xt}^{ucc+} \leq 0 \quad (365)$$

$$F_{xt}^{ucc+} = F_{xt}^{ucc} = F_{xt}^c \quad F_{xt}^{ucc+} \geq 0 \quad (366)$$

$$\beta_{v_x v_x}^{ucc+} = \beta_{v_x v_x}^{ucc} = -\pi/2 + \delta_{xt} \quad (367)$$

where  $E_{xt}^{ucc}$ ,  $F_{xt}^{ucc}$ ,  $E_{xt}^c$  and  $F_{xt}^c$  are given by equations (350), (351), (135) and (136) respectively, and where  $\delta_{xt}$  is given by equation (141). Equations (362) and (363) for repulsive forces can be rewritten to give Newton's dynamical law as

$$m_{xt}^{-2} [(E_{xt}^{ucc+})^2 + (F_{xt}^{ucc+})^2]^{1/2} = -W/x \partial \theta_W / \partial \theta_x \quad (368)$$

or equivalently

$$-2 [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} = -W/x \partial \theta_W / \partial \theta_x \quad (369)$$

where  $W \partial \theta_W / \partial \theta_x \leq 0$  for repulsive forces. The phase angle condition for Newton's law of motion equation (361) becomes for this special case

$$\theta_W = 2(\theta_x - \theta_t^x) + \delta_{xt} \quad (370)$$

The equivalent complex number form of Newton's law of motion for repulsive forces is given by

$$m\bar{x}/\bar{t}^2 (-E_{xt}^{ucc+} + jF_{xt}^{ucc+}) = -\bar{W}/\bar{x} \partial \theta_W / \partial \theta_x \quad (371)$$

or equivalently as

$$m\bar{x}/\bar{t}^2 (-E_{xt}^{ucc} + jF_{xt}^{ucc}) = -\bar{W}/\bar{x} \partial \theta_W / \partial \theta_x \quad (373)$$

$$m\bar{x}/\bar{t}^2 (-E_{xt}^c + jF_{xt}^c) = -\bar{W}/\bar{x} \partial \theta_W / \partial \theta_x \quad (374)$$

which are valid for a repulsive force with  $W \partial \theta_W / \partial \theta_x \leq 0$ .

#### E. Simple Harmonic Oscillator in Broken Symmetry Spacetime.

An elementary mechanical system that can be used to describe the effects of broken symmetry spacetime on a dynamical system is the simple harmonic oscillator.<sup>7-10</sup> The complex number potential energy for a simple harmonic oscillator in broken symmetry spacetime is given by

$$\bar{W} = 1/2 k \bar{x}^2 \quad W = 1/2 k x^2 \quad \theta_W = \theta_k + 2\theta_x \quad (374)$$

For the simple harmonic oscillator potential it follows from equations (4), (164) and (374) that

$$\beta_{WW} = \beta_{xx} \quad (375)$$

Because of the validity of equation (375) it follows for the simple harmonic oscillator that a slow mechanical process with  $\beta_{WW} = 0$  must necessarily occur in incoherent space with  $\beta_{xx} = 0$ , and an ultrafast mechanical process with  $\beta_{WW} = \pi/2$  must occur in coherent space with  $\beta_{xx} = \pi/2$ . The simple harmonic oscillator is an example of an attractive force system.

The equations of motion for the simple harmonic oscillator are now considered for four limiting spacetime conditions.

Case a. Incoherent Space and Incoherent Time (Slow Mechanical Process).

Equation (178) gives

$$d^2x/dt^2 + kx = 0 \quad (376)$$

which is the standard equation for the simple harmonic oscillator.

Case b. Coherent Space and Incoherent Time (Ultrafast Mechanical Process).

Equations (76), (290), (296) and (374) give

$$-m(E_{xt}^{ci} - jF_{xt}^{ci}) + \bar{k} = 0 \quad (377)$$

$$-m[(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} + k = 0 \quad (378)$$

$$\kappa_{xt} = \theta_k \quad (379)$$

where  $E_{xt}^{ci}$  and  $F_{xt}^{ci}$  are given by equations (69) and (70).

Case c. Incoherent Space and Coherent Time (Slow Mechanical Process).

Equations (107), (217), (222) and (223) give with  $\theta_x = 0$

$$-m/\bar{t}^2(F_{xt}^{ic} + jE_{xt}^{ic}) + \bar{k}x = 0 \quad (380)$$

$$-mt^{-2}[(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} + kx = 0 \quad (381)$$

$$\beta_{vxvx}^{ic} - 2\theta_t^x = \theta_k \quad (382)$$

where  $E_{xt}^{ic}$  and  $F_{xt}^{ic}$  are given by equations (103) and (104).

Case d. Coherent Space and Coherent Time (Ultrafast Mechanical Process).

Equations (146), (348), (353) and (374) give

$$- m/\bar{t}^2(-E_{xt}^c + jF_{xt}^c) + \bar{k} = 0 \quad (383)$$

$$- m\bar{t}^{-2}[(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} + k = 0 \quad (384)$$

$$\delta_{xt} - 2\theta_t^x = \theta_k \quad (385)$$

where  $E_{xt}^c$  and  $F_{xt}^c$  are given by equations (135) and (136). It is clear from cases b and d that the simple harmonic oscillator can undergo ultrafast internal spacetime motion in coherent space, and for these two cases the spring constant itself drives the internal motion.

**4. CONCLUSION.** Newton's law of dynamics can be developed for slow and ultrafast varying mechanical potential functions and for four possible variations of the spacetime coordinates: incoherent space and incoherent time, coherent space and incoherent time, incoherent space and coherent time, and coherent space and coherent time. This yields eight possible forms for Newton's law of motion. The ultrafast mechanical processes occur as rotations of the complex number potential function in an internal space while the magnitude of the complex number potential function remains fixed. The coherent changes of space and time coordinates occur as rotations of the complex number coordinates in an internal space for fixed magnitudes of the coordinates. Internal phase motions may possibly play an important role in the development of nuclear rocket engines.

#### ACKNOWLEDGEMENT

I would like to thank Elizabeth K. Klein for typing and editing this paper.

#### REFERENCES

1. Greiner, W., Müller, B., and Rafelski, J., Quantum Electrodynamics of Strong Fields, Springer-Verlag, New York, 1985.
2. Cheng, T. and Li, L., Gauge Theory of Elementary Particle Physics, Clarendon Press, Oxford, 1984.
3. Itzykson, C. and Zuber, J., Quantum Field Theory, McGraw-Hill, New York, 1980.
4. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
5. Weiss, R. A., Relativistic Thermodynamics, Vols. 1&2, K&W Publications, Vicksburg, MS, 1976.
6. Weiss, R. A., "Dynamical Systems in Asymmetric Space and Time," paper in Clean Fission, K&W Publications, Vicksburg, MS, 1992.
7. Corben, H. C. and Stehle, P., Classical Mechanics, Wiley, New York, 1957.
8. Lindsay, R. B., Physical Mechanics, Van Nostrand, New York, 1950.
9. Goldstein, H., Classical Mechanics, Addison-Wesley, Reading, MA, 1959.
10. Osgood, W. F., Mechanics, MacMillan, New York, 1937.

## ULTRAFAST QUANTUM PROCESSES

Richard A. Weiss  
U.S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** This paper develops the quantum mechanical theory of slow and ultrafast processes that occur in space and time with broken internal symmetries. For a slow quantum process the magnitude of the complex number wave function changes in space and time, while for an ultrafast process the complex number wave function rotates with a constant magnitude in an internal space. For incoherent space and time the changes in coordinates occur as the stretch or contraction of the magnitudes of the complex number spacetime coordinates, whereas for coherent spacetime the changes of the coordinates occur as rotations in an internal space. The case of a slow process in incoherent space is just the standard case of quantum mechanics. The Schrödinger and Dirac equations are developed for slow and ultrafast quantum processes in incoherent and coherent spacetime. For the case of an ultrafast process in coherent spacetime Schrödinger's equation describes the internal phase angle of the wave function in terms of the internal phase angles of the space and time coordinates. By way of example Schrödinger's equation for the ultrafast motion of a simple harmonic oscillator is developed for coherent spacetime.

**1. INTRODUCTION.** Ultrafast processes have become important diagnostic tools for the investigation of atomic and molecular phenomena that occur in gases and condensed matter, and the use of ultrashort picosecond ( $10^{-12}$  sec) and femtosecond ( $10^{-15}$  sec) laser light pulses has become the basic technique for studying ultrafast chemical and physical processes.<sup>1-17</sup> Physics, chemistry and biology have profited from ultrafast pulse technology because these light pulses can be used to examine atomic and molecular processes such as molecular vibrations in liquids, phonon and exciton decay in solids, time variation of plasma densities in gases and solids, chemical reactions at a molecular level and temperature fluctuation processes among many others.<sup>1-17</sup> In the visible region of the electromagnetic spectrum the Heisenberg uncertainty principle puts a limit of a femtosecond on the duration of a laser light pulse, while in the x ray region the limit of temporal resolution is an attosecond ( $10^{-18}$  sec) so that processes that occur faster by several orders of magnitude can in principle be observed as technology improves.<sup>1-17</sup>

Ultrafast processes require an appropriate thermodynamic description. A relativistic gauge theory of thermodynamics has been developed that determines the renormalized pressure and Grüneisen parameter in terms of the corresponding theoretically predicted unrenormalized values of these two quantities.<sup>18</sup> On the basis of this theory it was suggested that thermodynamic quantities such as internal energy, entropy and pressure are associated with broken internal symmetries and must be represented as complex numbers in an internal space.<sup>19</sup> Space and time coordinates also have broken internal symmetries and must likewise be represented by complex numbers in an internal space.<sup>19</sup> This is also true of kinematical and dynamical quantities such as velocity, momentum, acceleration and force.<sup>19,20</sup>

A theory of ultrafast processes has been developed that describes the

ultrafast variation of physical quantities as rotations in an internal space while the magnitudes of the physical quantities are fixed.<sup>19,20</sup> Thermodynamic quantities such as entropy, internal energy and volume are complex numbers that change as rotations in an internal space for ultrafast processes.<sup>19,20</sup> It has been suggested that thermodynamic engines can be developed whose cycles occur as changes in the internal phase angles of the entropy, internal energy and volume while the magnitudes of the quantities remain fixed.<sup>20</sup> Quantum mechanics can be developed with coordinates that have broken internal symmetries and the Schrödinger and Dirac equations have been formulated for partially coherent spacetime.<sup>19,20</sup> Applications to the problems of a particle confined to a box and the simple harmonic oscillator have been considered for the external motion in broken symmetry spacetime and for internal coordinate motion where the magnitudes of the coordinates are held fixed.<sup>20</sup>

For broken symmetry spacetime the space and time coordinates are written as complex numbers in an internal space in the following way

$$\bar{\alpha} = \alpha \exp(j\theta_{\alpha}) \quad \bar{t} = t \exp(j\theta_t) \quad (1)$$

where  $\alpha = x, y, z$  for cartesian coordinates.  $\alpha = r, \phi, z$  for cylindrical polar coordinates and  $\alpha = r, \phi, \psi$  for spherical polar coordinates. All physical quantities, with the exception of the light speed in the vacuum, have broken internal symmetries and must be represented as complex numbers in an internal space.<sup>19</sup> This includes, for example, pressure, entropy, energy and magnetic and electric field strengths. Therefore for the case of quantum mechanics the wave function must be represented as a complex number in internal space as follows<sup>21</sup>

$$\bar{\psi} = \psi \exp(j\theta_{\psi}) \quad (2)$$

Strictly speaking, the value of the internal phase angle of the time is associated with the particular physical quantity which is varying with time, so that for the case at hand<sup>21</sup>

$$\bar{t} = t \exp(j\theta_t^{\psi}) \quad (3)$$

where  $\theta_t^{\psi}$  = internal phase angle of time that is associated with the time variation of the wave function  $\bar{\psi}$ . Space is taken to be homogeneous so that the internal phase angle of time  $\theta_t^{\psi}$  is independent of the internal phase angles  $\theta_{\alpha}$  of the space coordinates. In other cases, such as particle dynamics and kinematics, the space and time coordinates are not independent and the internal phase angles of the space coordinates  $\theta_{\alpha}$  are each associated with a corresponding internal phase angle of time  $\theta_t^{\alpha}$  for  $\alpha = x, y, z$  with  $\partial\theta_{\alpha}/\partial\theta_t^{\alpha} \neq 0$ . But for the case of the wave equations of quantum mechanics the time and space coordinates are taken to be independent parameters so that  $\theta_t^{\psi}$  and  $\theta_{\alpha}$  are unrelated quantities with  $\partial\theta_{\alpha}/\partial\theta_t^{\psi} = 0$ . In general

$$\psi = \psi(\alpha, \theta_{\alpha}, t, \theta_t^{\psi}) \quad (4)$$

$$\theta_{\psi} = \theta_{\psi}(\alpha, \theta_{\alpha}, t, \theta_t^{\psi}) \quad (5)$$



where  $\alpha = x, y, z$ . From equations (1) and (2) it follows that<sup>19,21</sup>

$$d\bar{\alpha} = \sec \beta_{\alpha\alpha} d\alpha \exp[j(\theta_{\alpha} + \beta_{\alpha\alpha})] \quad (6)$$

$$= \csc \beta_{\alpha\alpha} \alpha d\theta_{\alpha} \exp[j(\theta_{\alpha} + \beta_{\alpha\alpha})] \quad (7)$$

$$d\bar{t} = \sec \beta_{tt}^{\Psi} dt \exp[j(\theta_t^{\Psi} + \beta_{tt}^{\Psi})] \quad (8)$$

$$= \csc \beta_{tt}^{\Psi} t d\theta_t^{\Psi} \exp[j(\theta_t^{\Psi} + \beta_{tt}^{\Psi})] \quad (9)$$

$$d\bar{\Psi} = \sec \beta_{\Psi\Psi} d\Psi \exp[j(\theta_{\Psi} + \beta_{\Psi\Psi})] \quad (10)$$

$$= \csc \beta_{\Psi\Psi} \Psi d\theta_{\Psi} \exp[j(\theta_{\Psi} + \beta_{\Psi\Psi})] \quad (11)$$

where

$$\tan \beta_{\alpha\alpha} = \alpha \partial \theta_{\alpha} / \partial \alpha \quad (12)$$

$$\tan \beta_{tt}^{\Psi} = t \partial \theta_t^{\Psi} / \partial t \quad (13)$$

$$\tan \beta_{\Psi\Psi} = \Psi \partial \theta_{\Psi} / \partial \Psi \quad (14)$$

The internal phase angle of the time  $\theta_t^{\Psi}$  is associated with the time variation of the wave function  $\Psi$ .

The first derivatives of the wave function with respect to space and time are written as

$$\bar{v}_{\alpha} = v_{\alpha} \exp(j\theta_{v\alpha}) = \partial \bar{\Psi} / \partial \bar{\alpha} \quad (15)$$

$$\bar{u} = u \exp(j\theta_u) = \partial \bar{\Psi} / \partial \bar{t} \quad (16)$$

where  $\alpha = x, y, z$ . For the space derivatives in equation (15) the magnitudes  $v_{\alpha}$  can be written as<sup>21</sup>

$$v_{\alpha} = \sec \beta_{\Psi\Psi} \cos \beta_{\alpha\alpha} \partial \Psi / \partial \alpha \quad (17)$$

$$= \csc \beta_{\Psi\Psi} \cos \beta_{\alpha\alpha} \Psi \partial \theta_{\Psi} / \partial \alpha \quad (18)$$

$$= \sec \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial \Psi / \partial \theta_{\alpha} \quad (19)$$

$$= \csc \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \Psi / \alpha \partial \theta_{\Psi} / \partial \theta_{\alpha} \quad (20)$$

and the internal phase angles as

$$\theta_{v\alpha} = \theta_{\Psi} + \beta_{\Psi\Psi} - \theta_{\alpha} - \beta_{\alpha\alpha} \quad (21)$$

where  $\beta_{\alpha\alpha}$  and  $\beta_{\psi\psi}$  are given by equations (12) and (14) respectively. The magnitude of the time derivative that appears in equation (16) is written as

$$u = \sec \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \partial\psi/\partial t \quad (22)$$

$$= \csc \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \psi \partial\theta_{\psi}/\partial t \quad (23)$$

$$= \sec \beta_{\psi\psi} \sin \beta_{tt}^{\psi} t^{-1} \partial\psi/\partial\theta_t^{\psi} \quad (24)$$

$$= \csc \beta_{\psi\psi} \sin \beta_{tt}^{\psi} \psi/t \partial\theta_{\psi}/\partial\theta_t^{\psi} \quad (25)$$

while the internal phase angle is given by

$$\theta_u = \theta_{\psi} + \beta_{\psi\psi} - \theta_t^{\psi} - \beta_{tt}^{\psi} \quad (26)$$

where  $\beta_{tt}^{\psi}$  is given by equation (13) and  $\theta_t^{\psi}$  = internal phase angle of time that is associated with the time variation of the wave amplitude  $\psi$ .

Four limiting forms of the first derivative of the wave function with respect to the spatial coordinates will now be considered.<sup>21</sup>

Case 1. Slow Quantum Process and Incoherent Space.

This is described by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (27)$$

Then equations (17) and (21) give

$$v_{\alpha}^{si} = \partial\psi/\partial\alpha \quad \theta_{v\alpha}^{si} = 0 \quad (28)$$

Case 2. Ultrafast Quantum Process and Incoherent Space.

This case is given by

$$\beta_{\psi\psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (29)$$

Equations (18) and (21) become for this case

$$v_{\alpha}^{ui} = \psi \partial\theta_{\psi}/\partial\alpha \quad \theta_{v\alpha}^{ui} = \theta_{\psi} + \pi/2 \quad (30)$$

or equivalently

$$\bar{v}_{\alpha}^{ui} = j\bar{\psi} \partial\theta_{\psi}/\partial\alpha \quad (31)$$

Case 3. Slow Quantum Process and Coherent Space.

The following conditions hold for this case

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad (32)$$

and equations (19) and (21) become

$$v_{\alpha}^{sc} = \alpha^{-1} \partial\Psi/\partial\theta_{\alpha} \quad \theta_{v\alpha}^{sc} = -\theta_{\alpha} - \pi/2 \quad (33)$$

or equivalently

$$\bar{v}_{\alpha}^{sc} = -j/\bar{\alpha} \partial\Psi/\partial\theta_{\alpha} \quad (34)$$

Case 4. Ultrafast Quantum Process and Coherent Space.

This case is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad (35)$$

and equations (20) and (21) give

$$v_{\alpha}^{uc} = \Psi/\alpha \partial\theta_{\Psi}/\partial\theta_{\alpha} \quad \theta_{v\alpha}^{uc} = \theta_{\Psi} - \theta_{\alpha} \quad (36)$$

or equivalently

$$\bar{v}_{\alpha}^{uc} = \bar{\Psi}/\bar{\alpha} \partial\theta_{\Psi}/\partial\theta_{\alpha} \quad (37)$$

These are the four limiting cases associated with the first derivative with respect to the spatial coordinates.<sup>21</sup>

Now the four limiting conditions will be given for the first derivative of the wave function with respect to time.<sup>21</sup>

Case 1. Slow Quantum Process and Incoherent Time.

This case is given by

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \theta_t^{\Psi} = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (38)$$

and equations (22) and (26) become

$$u^{si} = \partial\Psi/\partial t \quad \theta_u^{si} = 0 \quad (39)$$

Case 2. Ultrafast Quantum Process and Incoherent Time.

This case is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \theta_t^{\Psi} = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (40)$$

Equations (23) and (26) then give

$$u^{u1} = \psi \partial \theta_{\psi} / \partial t \quad \theta_u^{u1} = \theta_{\psi} + \pi/2 \quad (41)$$

or

$$\bar{u}^{u1} = j \bar{\psi} \partial \theta_{\psi} / \partial t \quad (42)$$

Case 3. Slow Quantum Process in Coherent Time.

The following conditions are valid for this case

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{tt}^{\psi} = \pi/2 \quad (43)$$

while equations (24) and (26) give

$$u^{sc} = t^{-1} \partial \psi / \partial \theta_t^{\psi} \quad \theta_u^{sc} = -\theta_t^{\psi} - \pi/2 \quad (44)$$

or

$$\bar{u}^{sc} = -j/\bar{t} \partial \psi / \partial \theta_t^{\psi} \quad (45)$$

Case 4. Ultrafast Quantum Process in Coherent Time.

This case is described by the following conditions

$$\beta_{\psi\psi} = \pi/2 \quad \beta_{tt}^{\psi} = \pi/2 \quad (46)$$

and equations (25) and (26) give

$$u^{uc} = \psi/t \partial \theta_{\psi} / \partial \theta_t^{\psi} \quad \theta_u^{uc} = \theta_{\psi} - \theta_t^{\psi} \quad (47)$$

which can be rewritten as

$$\bar{u}^{uc} = \bar{\psi}/\bar{t} \partial \theta_{\psi} / \partial \theta_t^{\psi} \quad (48)$$

The second derivatives of the wave function with respect to the spatial coordinates will now be represented in four general forms which can be specialized to four limiting cases of physical interest corresponding to slow and fast quantum processes in incoherent and coherent space. The second derivative of the wave function with respect to space coordinates is written as<sup>21</sup>

$$\bar{\xi}_{\alpha} = \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \partial^2 \bar{\psi} / \partial \bar{\alpha}^2 = \partial \bar{v}_{\alpha} / \partial \bar{\alpha} \quad (49)$$

where  $\alpha = x, y, z$ , and where  $\bar{v}_{\alpha}$  is defined in equation (15).

Case 1. Slow Quantum Process in Incoherent Space.

A general expression for the second spatial derivative of the wave func-

tion will be derived which can be used to pass to the limit of a slow quantum process in incoherent space which is described by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (50)$$

for  $\alpha = x, y$  and  $z$ . Equations (1), (2), (15), (17) and (49) give

$$\xi_{\alpha} = \sec \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} \partial v_{\alpha} / \partial \alpha \quad (51)$$

$$= \sec \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} \partial / \partial \alpha (\sec \beta_{\psi\psi} \cos \beta_{\alpha\alpha} \partial \Psi / \partial \alpha) \quad (52)$$

while equations (21) and (49) give

$$\theta_{\xi\alpha} = \theta_{v\alpha} + \beta_{v\alpha v\alpha} - \theta_{\alpha} - \beta_{\alpha\alpha} \quad (53)$$

$$= \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (54)$$

where  $\beta_{v\alpha v\alpha}$  is given by

$$\tan \beta_{v\alpha v\alpha} = v_{\alpha} \partial \theta_{v\alpha} / \partial v_{\alpha} \quad (55)$$

where  $v_{\alpha}$  is given by equation (17) and  $\theta_{v\alpha}$  by equation (21). For this case  $\theta_{\xi\alpha}$  is a small number. In the limiting case of a slow quantum process in incoherent space equation (50) is valid and equations (52) and (54) become

$$\xi_{\alpha}^{si} = \partial^2 \Psi / \partial \alpha^2 \quad \theta_{\xi\alpha}^{si} = 0 \quad (56)$$

which is the conventional result.

## Case 2. Ultrafast Quantum Process in Incoherent Space.

This section derives a general equation for the second derivatives of the wave function with respect to the space coordinates which can be utilized to obtain the limiting case of an ultrafast quantum process in incoherent space which is defined by

$$\beta_{\psi\psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (57)$$

where  $\theta_{\psi}$  is now a variable. From equations (18) and (49) it follows that

$$\xi_{\alpha} = \csc \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} v_{\alpha} \partial \theta_{v\alpha} / \partial \alpha \quad (58)$$

$$= \csc \beta_{v\alpha v\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \Psi \partial \theta_{\psi} / \partial \alpha \partial \theta_{v\alpha} / \partial \alpha \quad (59)$$

where  $\beta_{v\alpha v\alpha}$  is given by equation (55) with  $v_{\alpha}$  given by equation (18) and where equation (21) gives

$$\partial \theta_{v\alpha} / \partial \alpha = \partial / \partial \alpha (\theta_{\psi} + \beta_{\psi\psi} - \theta_{\alpha} - \beta_{\alpha\alpha}) \quad (60)$$

The corresponding internal phase angle for the second spatial derivative is given by

$$\theta_{\xi\alpha} = \theta_{v\alpha} + \beta_{v\alpha v\alpha} - \theta_{\alpha} - \beta_{\alpha\alpha} \quad (61)$$

$$= \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (62)$$

For an ultrafast quantum process it is convenient to introduce an alternative representation of the second derivatives which is given by

$$\bar{\xi}_{\alpha} = \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \xi_{\alpha}^{\dagger} \exp(j\theta_{\xi\alpha}^{\dagger}) \quad (63)$$

$$= \partial^2 \bar{\psi} / \partial \bar{\alpha}^2 = \partial \bar{v}_{\alpha} / \partial \bar{\alpha}$$

where

$$\xi_{\alpha}^{\dagger} = -\xi_{\alpha} \quad (64)$$

$$\theta_{\xi\alpha}^{\dagger} = \theta_{\xi\alpha} - \pi \quad (65)$$

Then it follows that

$$\xi_{\alpha}^{\dagger} = -\csc \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} v_{\alpha} \partial \theta_{v\alpha} / \partial \alpha \quad (66)$$

$$= -\csc \beta_{v\alpha v\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi \partial \theta_{\psi} / \partial \alpha \partial \theta_{v\alpha} / \partial \alpha \quad (67)$$

and

$$\theta_{\xi\alpha}^{\dagger} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) - \pi \quad (68)$$

for the general case.

In the limiting case of an ultrafast quantum process in incoherent space, equation (57) is valid and equations (54) and (59) become

$$\xi_{\alpha}^{ui} = \csc \beta_{v\alpha v\alpha}^{ui} \psi (\partial \theta_{\psi} / \partial \alpha)^2 \quad (69)$$

$$\theta_{\xi\alpha}^{ui} = \theta_{\psi} + \beta_{v\alpha v\alpha}^{ui} + \pi/2 \quad (70)$$

From equations (30) and (55) it follows that

$$\tan \beta_{v\alpha v\alpha}^{ui} = E_{\psi\alpha}^{ui} / F_{\psi\alpha}^{ui} \quad (71)$$

where

$$E_{\psi\alpha}^{ui} = (\partial \theta_{\psi} / \partial \alpha)^2 \quad E_{\psi\alpha}^{ui} \geq 0 \quad (72)$$

$$F_{\psi\alpha}^{ui} = \partial^2 \theta_{\psi} / \partial \alpha^2 \quad F_{\psi\alpha}^{ui} \leq 0 \quad (73)$$

From equation (71) it follows that

$$\csc \beta_{\psi\alpha}^{ui} = [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} / E_{\psi\alpha}^{ui} \quad (74)$$

Equations (69) and (74) give

$$\xi_{\alpha}^{ui} = \psi [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} \quad (75)$$

For the signs chosen in equations (72) and (73) it follows from equation (71) that

$$\beta_{\psi\alpha}^{ui} = \pi/2 + \kappa_{\psi\alpha} \quad (76)$$

where  $\kappa_{\psi\alpha}$  is a small positive number defined by

$$\tan \kappa_{\psi\alpha} = |F_{\psi\alpha}^{ui}| / E_{\psi\alpha}^{ui} \quad (77)$$

Equations (70) and (76) give

$$\theta_{\xi\alpha}^{ui} = \theta_{\psi} + \kappa_{\psi\alpha} + \pi \quad (78)$$

and equations (64) and (65) give

$$\xi_{\alpha}^{ui+} = -\psi [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} \quad (79)$$

$$\theta_{\xi\alpha}^{ui+} = \theta_{\psi} + \kappa_{\psi\alpha} \quad (80)$$

so that  $\theta_{\xi\alpha}^{ui+}$  is a small number. Finally, from equation (49) and (69) through (80) it follows that

$$\bar{\xi}_{\alpha}^{ui} = -\bar{\psi}(E_{\psi\alpha}^{ui} - jF_{\psi\alpha}^{ui}) \quad (81)$$

which is the equivalent complex number representation of equations (79) and (80).

### Case 3. Slow Quantum Process in Coherent Space.

An expression is derived for the second derivative of the wave function with respect to the spatial coordinates, which can be used to pass to the limit of a slow quantum process in coherent space whose characteristics are

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad (82)$$

where  $\theta_{\alpha}$  is variable. From equations (49), (19) and (54) it follows that

$$\xi_{\alpha} = \sec \beta_{\psi\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial v_{\alpha} / \partial \theta_{\alpha} \quad (83)$$

$$= \sec \beta_{\psi\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial / \partial \theta_{\alpha} (\sec \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial \psi / \partial \theta_{\alpha}) \quad (84)$$

$$\theta_{\xi\alpha} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (85)$$

where  $\beta_{v\alpha v\alpha}$  is given by (19), (21) and (55). In this case it is convenient to introduce another representation of the second spatial derivative, namely

$$\begin{aligned} \bar{\xi}_{\alpha} &= \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \xi'_{\alpha} \exp(j\theta'_{\xi\alpha}) \\ &= \partial^2 \bar{\psi} / \partial \bar{\alpha}^2 = \partial \bar{v}_{\alpha} / \partial \bar{\alpha} \end{aligned} \quad (86)$$

where

$$\xi'_{\alpha} = -\xi_{\alpha} \quad (87)$$

$$\theta'_{\xi\alpha} = \theta_{\xi\alpha} + \pi \quad (88)$$

The limiting case of a slow quantum process in coherent space is obtained from equation (82) which combined with equations (84) and (85) gives

$$\xi_{\alpha}^{sc} = \sec \beta_{v\alpha v\alpha}^{sc} \alpha^{-2} \partial^2 \psi / \partial \theta_{\alpha}^2 \quad (89)$$

$$\theta_{\xi\alpha}^{sc} = \beta_{v\alpha v\alpha}^{sc} - 2\theta_{\alpha} - \pi \quad (90)$$

Equations (33) and (55) give

$$\tan \beta_{v\alpha v\alpha}^{sc} = E_{\psi\alpha}^{sc} / F_{\psi\alpha}^{sc} \quad (91)$$

where

$$E_{\psi\alpha}^{sc} = -\partial \psi / \partial \theta_{\alpha} \quad E_{\psi\alpha}^{sc} \geq 0 \quad (92)$$

$$F_{\psi\alpha}^{sc} = \partial^2 \psi / \partial \theta_{\alpha}^2 \quad F_{\psi\alpha}^{sc} \geq 0 \quad (93)$$

$$\sec \beta_{v\alpha v\alpha}^{sc} = [(E_{\psi\alpha}^{sc})^2 + (F_{\psi\alpha}^{sc})^2]^{1/2} / F_{\psi\alpha}^{sc} \quad (94)$$

and therefore  $\beta_{v\alpha v\alpha}^{sc}$  is a small positive angle. Equations (89), (92) and (94) give

$$\xi_{\alpha}^{sc} = \alpha^{-2} [(E_{\psi\alpha}^{sc})^2 + (F_{\psi\alpha}^{sc})^2]^{1/2} \quad (95)$$

The alternative description of the second derivative with respect to space is obtained from equations (87), (88), (90) and (95) as

$$\xi_{\alpha}^{sc'} = -\alpha^{-2} [(E_{\psi\alpha}^{sc})^2 + (F_{\psi\alpha}^{sc})^2]^{1/2} \quad (96)$$

$$\theta_{\xi\alpha}^{sc'} = \beta_{v\alpha v\alpha}^{sc} - 2\theta_{\alpha} \quad (97)$$



where  $\theta_{\xi\alpha}^{sc}$  is seen to be a small angle. The complex number that is equivalent to equations (49), (86), (96) and (97) is

$$\bar{\xi}_{\alpha}^{sc} = -1/\bar{\alpha}^2 (F_{\Psi\alpha}^{sc} + jE_{\Psi\alpha}^{sc}) \quad (98)$$

which gives the second order spatial derivatives of the wave function for a slow quantum process in coherent space.

#### Case 4. Ultrafast Quantum Process and Coherent Space.

This section develops a representation for the second derivative of the wave function with respect to the space coordinates which can be utilized to attain the limit of an ultrafast quantum process in coherent space whose description is

$$\beta_{\Psi\Psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad (99)$$

The magnitude and internal phase angle of the second derivative is obtained from equations (20), (49) and (54) to be

$$\xi_{\alpha} = \csc \beta_{\nu\alpha\nu\alpha} \sin \beta_{\alpha\alpha} \nu_{\alpha}/\alpha \partial\theta_{\nu\alpha}/\partial\theta_{\alpha} \quad (100)$$

$$= \csc \beta_{\nu\alpha\nu\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{\Psi\Psi} \Psi/\alpha^2 \partial\theta_{\Psi}/\partial\theta_{\alpha} \partial\theta_{\nu\alpha}/\partial\theta_{\alpha} \quad (101)$$

$$\theta_{\xi\alpha} = \theta_{\Psi} + \beta_{\Psi\Psi} + \beta_{\nu\alpha\nu\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (102)$$

where  $\beta_{\nu\alpha\nu\alpha}$  is given by equations (20) and (55), and where from equation (31) it follows that

$$\partial\theta_{\nu\alpha}/\partial\theta_{\alpha} = \partial\theta_{\Psi}/\partial\theta_{\alpha} - 1 + \partial/\partial\theta_{\alpha} (\beta_{\Psi\Psi} - \beta_{\alpha\alpha}) \quad (103)$$

A different expression for the second derivative can be obtained from equations (83) and (20) and is

$$\xi_{\alpha} = \sec \beta_{\nu\alpha\nu\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial\nu_{\alpha}/\partial\theta_{\alpha} \quad (104)$$

$$= \sec \beta_{\nu\alpha\nu\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\csc \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \Psi/\alpha \partial\theta_{\Psi}/\partial\theta_{\alpha}) \quad (105)$$

A comparison of equations (101) and (105) gives

$$\tan \beta_{\nu\alpha\nu\alpha} = C_{\Psi\alpha}/D_{\Psi\alpha} \quad (106)$$

$$\csc \beta_{\nu\alpha\nu\alpha} = (C_{\Psi\alpha}^2 + D_{\Psi\alpha}^2)^{1/2}/C_{\Psi\alpha} \quad (107)$$

$$\sec \beta_{\nu\alpha\nu\alpha} = (C_{\Psi\alpha}^2 + D_{\Psi\alpha}^2)^{1/2}/D_{\Psi\alpha} \quad (108)$$

where

$$C_{\Psi\alpha} = \sin \beta_{\alpha\alpha} \csc \beta_{\Psi\Psi} \Psi/\alpha \partial\theta_{\Psi}/\partial\theta_{\alpha} \partial\theta_{\Psi\alpha}/\partial\theta_{\alpha} \quad (109)$$

$$D_{\Psi\alpha} = \partial/\partial\theta_{\alpha} (\csc \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \Psi/\alpha \partial\theta_{\Psi}/\partial\theta_{\alpha}) \quad (110)$$

where  $\partial\theta_{\Psi\alpha}/\partial\theta_{\alpha}$  is given by equation (103) and where  $C_{\Psi\alpha} \leq 0$ . Therefore,

$$\xi_{\alpha} = C_{\Psi\alpha} \alpha^{-1} \csc \beta_{\Psi\alpha\Psi} \sin \beta_{\alpha\alpha} \quad (111)$$

$$= D_{\Psi\alpha} \alpha^{-1} \sec \beta_{\Psi\alpha\Psi} \sin \beta_{\alpha\alpha} \quad (112)$$

which can be rewritten as

$$\xi_{\alpha} = (C_{\Psi\alpha}^2 + D_{\Psi\alpha}^2)^{1/2} \alpha^{-1} \sin \beta_{\alpha\alpha} \quad (113)$$

It is convenient for the case at hand to write the second derivative of the wave function with respect to space coordinates in equation (49) as

$$\begin{aligned} \bar{\xi}_{\alpha} &= \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \xi'_{\alpha} \exp(j\theta'_{\xi\alpha}) \\ &= \partial^2 \bar{\Psi} / \partial \bar{\alpha}^2 = \partial \bar{V}_{\alpha} / \partial \bar{\alpha} \end{aligned} \quad (114)$$

where

$$\xi'_{\alpha} = -\xi_{\alpha} \quad (115)$$

$$\theta'_{\xi\alpha} = \theta_{\xi\alpha} + \pi \quad (116)$$

which gives a useful alternative description of the second derivative with respect to space.

For an ultrafast quantum process in coherent space described by equation (99) it follows from equations (101), (105), (113) and (115) that

$$\xi_{\alpha}^{uc} = -\csc \beta_{\Psi\alpha\Psi}^{uc} \Psi/\alpha^2 E_{\Psi\alpha}^{uc} \quad (117)$$

$$= -\sec \beta_{\Psi\alpha\Psi}^{uc} \Psi/\alpha^2 F_{\Psi\alpha}^{uc} \quad (118)$$

$$= -\alpha^{-1} [(C_{\Psi\alpha}^{uc})^2 + (D_{\Psi\alpha}^{uc})^2]^{1/2} \quad (119)$$

$$= -\Psi\alpha^{-2} [(E_{\Psi\alpha}^{uc})^2 + (F_{\Psi\alpha}^{uc})^2]^{1/2} \quad (120)$$

where

$$E_{\Psi\alpha}^{uc} = \partial\theta_{\Psi}/\partial\theta_{\alpha} (\partial\theta_{\Psi}/\partial\theta_{\alpha} - 1) \quad E_{\Psi\alpha}^{uc} \leq 0 \quad (121)$$

$$F_{\Psi\alpha}^{uc} = \partial^2 \theta_{\Psi} / \partial \theta_{\alpha}^2 \quad F_{\Psi\alpha}^{uc} \geq 0 \quad (122)$$

$$C_{\Psi\alpha}^{uc} = \Psi/\alpha E_{\Psi\alpha}^{uc} \quad (123)$$

$$D_{\Psi\alpha}^{uc} = \Psi/\alpha F_{\Psi\alpha}^{uc} \quad (124)$$

From equations (36) and (55) or directly from equation (106) it follows that

$$\tan \beta_{\Psi\alpha}^{uc} = E_{\Psi\alpha}^{uc}/F_{\Psi\alpha}^{uc} \quad (125)$$

$$\csc \beta_{\Psi\alpha}^{uc} = [(E_{\Psi\alpha}^{uc})^2 + (F_{\Psi\alpha}^{uc})^2]^{1/2}/E_{\Psi\alpha}^{uc} \quad (126)$$

$$\sec \beta_{\Psi\alpha}^{uc} = [(E_{\Psi\alpha}^{uc})^2 + (F_{\Psi\alpha}^{uc})^2]^{1/2}/F_{\Psi\alpha}^{uc} \quad (127)$$

Because of the choice of signs in equations (121) and (122) it follows that

$$\beta_{\Psi\alpha}^{uc} = -\pi/2 + \delta_{\Psi\alpha} \quad (128)$$

where  $\delta_{\Psi\alpha} > 0$ , so that

$$\tan \delta_{\Psi\alpha} = F_{\Psi\alpha}^{uc}/|E_{\Psi\alpha}^{uc}| \quad (129)$$

The internal phase angle of the second derivative then follows from equations (99), (102), (116) and (128) as

$$\theta_{\xi\alpha}^{uc} = \theta_{\Psi} + \beta_{\Psi\alpha}^{uc} - 2\theta_{\alpha} - \pi/2 \quad (130)$$

$$= \theta_{\Psi} - 2\theta_{\alpha} + \delta_{\Psi\alpha} - \pi \quad (131)$$

$$\theta_{\xi\alpha}^{uc'} = \theta_{\Psi} + \beta_{\Psi\alpha}^{uc} - 2\theta_{\alpha} + \pi/2 \quad (132)$$

$$= \theta_{\Psi} - 2\theta_{\alpha} + \delta_{\Psi\alpha} \quad (133)$$

so that  $\theta_{\xi\alpha}^{uc'}$  is a small number. The complex number second derivative with respect to space coordinates that corresponds to equations (120) and (133) is given by

$$\bar{\xi}_{\alpha}^{uc} = (\partial^2 \bar{\Psi} / \partial \bar{\alpha}^2)^{uc} = \bar{\Psi} / \bar{\alpha}^2 [\partial \theta_{\Psi} / \partial \theta_{\alpha} (\partial \theta_{\Psi} / \partial \theta_{\alpha} - 1) - j \partial^2 \theta_{\Psi} / \partial \theta_{\alpha}^2] \quad (134)$$

$$= -\bar{\Psi} / \bar{\alpha}^2 (-E_{\Psi\alpha}^{uc} + j F_{\Psi\alpha}^{uc}) \quad (135)$$

for an ultrafast quantum process in coherent spacetime.

In terms of the elementary expressions for the space and time derivatives of the wave function it is possible to formulate the quantum mechanical theory of ultrafast processes in spacetime with broken internal symmetries. Briefly the outline of this paper is as follows: Section 2 examines the time independent Schrödinger equation for slow and ultrafast quantum processes in cartesian

space with broken internal symmetries, Section 3 investigates the time dependent Schrödinger equation, and Section 4 considers the slow and fast forms of the Dirac equation in asymmetric spacetime.

**2. TIME INDEPENDENT SCHRÖDINGER EQUATION IN BROKEN SYMMETRY CARTESIAN COORDINATES.** This section develops the time independent Schrödinger equation for slow (incoherent) and ultrafast (coherent) spatial variations of the quantum wave function in cartesian space with broken internal symmetries. The space coordinates also can vary incoherently or coherently. The incoherent variation of the space coordinates corresponds to changes in the magnitudes of the coordinates, while the coherent variation of the space coordinates corresponds to rotations of the complex number coordinates in an internal space. For a slow (incoherent) quantum process the wave function is a scalar in internal space and changes in magnitude, while for an ultrafast (coherent) quantum process the complex number wave function rotates in an internal space. The standard forms of the equations of quantum mechanics correspond to an incoherent process in incoherent space.<sup>22-26</sup> Quantum mechanics with complex number wave functions in partially coherent spacetime has already appeared in the literature.<sup>19,20</sup> Various forms of Schrödinger's equation have been written for coherent spacetime where the change in the complex number coordinates is in the form of a rotation in internal space.<sup>20</sup> This section develops Schrödinger's time independent equation for slow and ultrafast quantum processes in incoherent and coherent space, so that four special cases are considered: slow quantum process in incoherent space, slow quantum process in coherent space, ultrafast quantum process in incoherent space, and an ultrafast quantum process in coherent space. A slow quantum process is assumed to have a wave function that varies incoherently in space, while an ultrafast process is assumed to have a wave function that varies coherently in space.

#### A. Schrödinger's Equation and Linear Momentum in Spacetime with Broken Internal Symmetries.

Schrödinger's equation for a slow process in incoherent space is treated fully in the literature.<sup>22-26</sup> For the general case of an asymmetric wave function in asymmetric spacetime Schrödinger's time dependent equation is written as<sup>27</sup>

$$-\hbar^2/(2\mu)\bar{\nabla}^2\bar{\psi} + \bar{V}\bar{\psi} = i\hbar\partial\bar{\psi}/\partial\bar{t} \quad (136)$$

where  $\hbar = h/(2\pi)$ ,  $h$  = Planck's constant,  $\mu$  = mass of particle,  $\bar{\nabla}^2$  = Laplacian expressed in terms of complex number coordinates,  $\bar{\psi}$  = complex number wave function which is represented by equation (2),  $\bar{V}$  = complex number potential which is written as<sup>27</sup>

$$\bar{V} = V \exp(j\theta_V) \quad (137)$$

and where  $\bar{t}$  = complex number time as described by equation (1) or more carefully by equation (3).

If the wave function is written in a form that is suitable for a stationary state as<sup>27</sup>

$$\bar{\Psi} = \bar{U}(\alpha) \exp(-i\bar{E}t/\hbar) \quad (138)$$

then equation (136) reduces to the following complex number time independent Schrodinger equation

$$-\hbar^2/(2\mu)\bar{\nabla}^2\bar{U} + \bar{V}\bar{U} = \bar{E}\bar{U} \quad (139)$$

where the complex number energy  $\bar{E}$  and wave function  $\bar{U}$  are written as

$$\bar{E} = E \exp(j\theta_E) \quad \bar{U} = U \exp(j\theta_U) \quad (140)$$

Schrödinger's equation (139) is obtained from the law of the conservation of energy<sup>27</sup>

$$[\sum_{\alpha} \bar{p}_{\alpha}^2/(2\mu) + \bar{V}]\bar{U} = \bar{E}\bar{U} \quad (141)$$

combined with the following momentum operator representation for complex number cartesian coordinates<sup>27</sup>

$$\bar{p}_{\alpha} = -i\hbar\partial/\partial\bar{\alpha} \quad (142)$$

$$\bar{p}_{\alpha}\bar{U} = -i\hbar\partial\bar{U}/\partial\bar{\alpha} = -i\hbar\bar{w}_{\alpha} \quad (143)$$

$$\bar{w}_{\alpha} = w_{\alpha} \exp(j\theta_{w\alpha}) = \partial\bar{U}/\partial\bar{\alpha} \quad (144)$$

where  $w_{\alpha}$  and  $\theta_{w\alpha}$  are given by

$$w_{\alpha} = \sec \beta_{UU} \cos \beta_{\alpha\alpha} \partial U/\partial\alpha \quad (145)$$

$$= \csc \beta_{UU} \cos \beta_{\alpha\alpha} U \partial\theta_U/\partial\alpha \quad (146)$$

$$= \sec \beta_{UU} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial U/\partial\theta_{\alpha} \quad (147)$$

$$= \csc \beta_{UU} \sin \beta_{\alpha\alpha} U/\alpha \partial\theta_U/\partial\theta_{\alpha} \quad (148)$$

$$\theta_{w\alpha} = \theta_U + \beta_{UU} - \theta_{\alpha} - \beta_{\alpha\alpha} \quad (149)$$

For a slow process in incoherent space equations (17) and (143) give

$$p_{\alpha}^{si}U = -i\hbar\partial U/\partial\alpha \quad (150)$$

$$(p_{\alpha}^{si})^2U = -\hbar^2\partial^2U/\partial\alpha^2 \quad (151)$$

which is the standard result. For an arbitrary variation of the complex number

wave function and for coherent space coordinate variations the momentum operator equations (142) and (143) become<sup>27</sup>

$$\bar{p}_\alpha = i\hbar/\bar{\alpha} \partial/\partial\theta_\alpha \quad \bar{p}_\alpha \bar{U} = i\hbar/\bar{\alpha} \partial\bar{U}/\partial\theta_\alpha \quad (152)$$

Applying the momentum operators in equations (142) and (152) twice gives the following results for coherent space variations<sup>27</sup>

$$\bar{p}_\alpha^2 = -\hbar^2 \partial^2/\partial\bar{\alpha}^2 = \hbar^2/\bar{\alpha}^2 (\partial^2/\partial\theta_\alpha^2 - j\partial/\partial\theta_\alpha) \quad (153)$$

$$\bar{p}_\alpha^2 \bar{U} = -\hbar^2 \partial^2 \bar{U}/\partial\bar{\alpha}^2 \quad (154)$$

$$= \hbar^2/\bar{\alpha}^2 (\partial^2 \bar{U}/\partial\theta_\alpha^2 - j\partial\bar{U}/\partial\theta_\alpha) \quad (155)$$

Equation (155) will subsequently be specialized to the case of an ultrafast quantum process where  $\bar{U}$  changes coherently.

Schrödinger's equation (139) can be written as

$$\bar{V}^2 \bar{U} + \bar{k}^2 \bar{U} = 0 \quad (156)$$

where

$$\bar{k}^2 = 2\mu/\hbar^2 (\bar{E} - \bar{V}) \quad (157)$$

with

$$\bar{k} = k \exp(j\theta_k) \quad (158)$$

The components of equation (157) are written as

$$k^2 \cos(2\theta_k) = C \quad (159)$$

$$k^2 \sin(2\theta_k) = D \quad (160)$$

where

$$C = 2\mu/\hbar^2 (E \cos \theta_E - V \cos \theta_V) \quad (161)$$

$$D = 2\mu/\hbar^2 (E \sin \theta_E - V \sin \theta_V) \quad (162)$$

so that  $k$  and  $\theta_k$  are given by

$$\tan 2\theta_k = D/C \quad (163)$$

$$k = (C^2 + D^2)^{1/4} \quad (164)$$

A crude approximation derived from equation (157) is

$$k^2 \sim k_s^2 = 2\mu/\hbar^2(E - V) \quad (165)$$

$$2\theta_k \sim 2\theta_{ks} = \theta_E = \theta_V \quad (166)$$

which assumes that the internal phase angles of the component terms in equation (157) are equal.

For cartesian coordinates Schrödinger's equation (156) is written as

$$-(\partial^2/\partial \bar{x}^2 + \partial^2/\partial \bar{y}^2 + \partial^2/\partial \bar{z}^2)\bar{U} = \bar{k}^2\bar{U} \quad (167)$$

where  $\bar{x}$ ,  $\bar{y}$  and  $\bar{z}$  are complex number cartesian coordinates that are represented by equation (1). Combining equations (49) and (167) gives Schrödinger's equation as

$$-\sum_{\alpha} \bar{\eta}_{\alpha} = \bar{k}^2\bar{U} \quad (168)$$

where

$$\bar{\eta}_{\alpha} = \eta_{\alpha} \exp(j\theta_{\eta\alpha}) = \partial^2\bar{U}/\partial \bar{\alpha}^2 = \partial \bar{w}_{\alpha}/\partial \bar{\alpha} \quad (169)$$

The complex number second derivative  $\bar{\eta}_{\alpha}$  is homologous to the complex number second derivative  $\bar{\xi}_{\alpha}$  of equations (49) through (135). All of the relations that appear in equations (49) through (135) involving  $\xi_{\alpha}$  and  $\theta_{\xi\alpha}$  have analogous expressions with  $\eta_{\alpha}$  and  $\theta_{\eta\alpha}$  and correspond to the replacement  $\bar{\Psi} \rightarrow \bar{U}$  and  $\bar{v}_{\alpha} \rightarrow \bar{w}_{\alpha}$ . Equation (168) can be approximated by assuming that the internal phase angles of each term in equation (168) are equal. Because the internal phase angles of the second derivative  $\bar{\eta}_{\alpha}$  can have a zero or a  $\pm \pi$  added as shown in Section 1 in equations (54), (65), (88) and (116), the approximate solution to equation (168) requires that the four possible states associated with a slow or ultrafast quantum process in incoherent or coherent space be considered separately. Schrödinger's equation (168) can be written approximately as<sup>27</sup>

$$-(\pm 1)\sum_{\alpha} \eta_{\alpha} = k^2 U \quad (170)$$

$$\theta_{\eta\alpha} \pm (\pi, 0) = 2\theta_k + \theta_U \quad (171)$$

where  $\theta_k$  and  $k$  are given by equations (163) and (164), and where  $\theta_{\eta\alpha}$  is given by equations (54) and (55) as

$$\theta_{\eta\alpha} = \theta_U + \beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (172)$$

where in a fashion similar to equations (14) and (55)

$$\tan \beta_{UU} = U\partial\theta_U/\partial U \quad \tan \beta_{w\alpha w\alpha} = w_{\alpha}\partial\theta_{w\alpha}/\partial w_{\alpha} \quad (173)$$

From equation (54) it follows that equation (171) can be written as

$$\beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \pm (\pi, 0) = 2\theta_k \quad (174)$$

The minus sign in the parenthesis in equation (170) appears whenever the  $\pm \pi$  appears in the phase angle conditions given by equations (171) or (174). Further approximations to Schrödinger's equations (170) through (174) can be obtained by writing

$$- (\pm 1) \sum_{\alpha} \eta_{\alpha} = k_s^2 U \quad (175)$$

$$\theta_{\eta\alpha} \pm (\pi, 0) = 2\theta_{ks} + \theta_U \quad (176)$$

$$\beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \pm (\pi, 0) = 2\theta_{ks} \quad (177)$$

where  $k_s$  and  $\theta_{ks}$  are given by equations (165) and (166).

#### B. Four Possible Cases of Quantum Processes in Space with Broken Internal Symmetries.

The four possible states associated with the time independent Schrödinger equations (170) and (171) that correspond to slow and ultrafast quantum processes and to incoherent and coherent spatial coordinate variations will now be considered.

##### Case a. Slow Quantum Processes in Incoherent Space.

For this case the momentum equations (143), (145) and (149) become

$$\bar{p}_{\alpha} \bar{U} = -i\hbar \sec \beta_{UU} \cos \beta_{c\alpha} \partial U / \partial \alpha \exp(j\theta_{w\alpha}) \quad (178)$$

where  $\theta_{w\alpha}$  is given by equation (149). For this case the cartesian form of Schrödinger's equation (167) combined with equations (51) through (54) gives

$$- \sum_{\alpha} \sec \beta_{w\alpha w\alpha} \cos \beta_{\alpha\alpha} \partial / \partial \alpha (\sec \beta_{UU} \cos \beta_{\alpha\alpha} \partial U / \partial \alpha) \exp(j\theta_{\eta\alpha}) = \bar{k}^2 \bar{U} \quad (179)$$

where  $\theta_{\eta\alpha}$  is given by equation (172). Equation (179) can be written approximately by assuming that the internal phase angles of the component terms are equal, which gives Schrödinger's equation as

$$- \sum_{\alpha} \sec \beta_{w\alpha w\alpha} \cos \beta_{\alpha\alpha} \partial / \partial \alpha (\sec \beta_{UU} \cos \beta_{\alpha\alpha} \partial U / \partial \alpha) = k^2 U \quad (180)$$

$$\theta_{\eta\alpha} = 2\theta_k + \theta_U \quad (181)$$

where equation (181) is valid for  $\alpha = x, y$  and  $z$ . Combining equations (172) and (181) gives



$$\beta_{UU} + \beta_{waw\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = 2\theta_k \quad (182)$$

In a further approximation Schrödinger's equation (179) can be written analogously to equations (180) and (181) using equations (165) and (166) as

$$- \sum_{\alpha} \sec \beta_{waw\alpha} \cos \beta_{\alpha\alpha} \partial/\partial\alpha (\sec \beta_{UU} \cos \beta_{\alpha\alpha} \partial U/\partial\alpha) = k_s^2 U \quad (183)$$

$$\theta_{\eta\alpha} = 2\theta_{ks} + \theta_U = \theta_V + \theta_U = \theta_E + \theta_U \quad (184)$$

Equation (184) is equivalent to

$$\beta_{UU} + \beta_{waw\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = 2\theta_{ks} = \theta_V = \theta_E \quad (185)$$

When all internal phase angles are set equal to zero equations (178) and (179) become the following standard equations of quantum mechanics for slow processes in incoherent space

$$p_{\alpha}^{si} U = -i\hbar \partial U/\partial\alpha \quad (186)$$

$$- \sum_{\alpha} \partial^2 U/\partial\alpha^2 = k_s^2 U \quad (187)$$

where  $k_s$  is given by equation (165).

Case b. Slow Quantum Processes in Coherent Space.

In this case the momentum equations (143), (147) and (149) become

$$\bar{p}_{\alpha} \bar{U} = -i\hbar \sec \beta_{UU} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial U/\partial\theta_{\alpha} \exp(j\theta_{w\alpha}) \quad (188)$$

which can be used to obtain the limiting case of a slow quantum process in coherent space. The general form of Schrödinger's equation (168) combined with equations (84) and (85) gives for this case

$$\begin{aligned} & - \sum_{\alpha} \sec \beta_{waw\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\sec \beta_{UU} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial U/\partial\theta_{\alpha}) \exp(j\theta_{\eta\alpha}) \\ & = \bar{k}^2 \bar{U} \end{aligned} \quad (189)$$

where  $\theta_{\eta\alpha}$  is given by equation (172). An approximate representation of equation (189) is obtained by assuming that the phase angles of the component terms of this equation are all equal and using the representation in equations (87) and (88) with the result that Schrödinger's equation is written as

$$\sum_{\alpha} \sec \beta_{waw\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\sec \beta_{UU} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial U/\partial\theta_{\alpha}) = k^2 U \quad (190)$$

$$\theta'_{\eta\alpha} = 2\theta_k + \theta_U \quad (191)$$

Equation (191) can be rewritten with the help of equations (85) and (88) as

$$\beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = 2\theta_k \quad (192)$$

where  $\theta_k$  and  $k$  are given by equations (163) and (164). A further approximate representation of Schrödinger's equation (189) is obtained by using equations (159) and (160) as

$$\sum_{\alpha} \sec \beta_{w\alpha w\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial \theta_{\alpha} (\sec \beta_{UU} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial U/\partial \theta_{\alpha}) = k_s^2 U \quad (193)$$

$$\beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = 2\theta_{ks} = \theta_E = \theta_V \quad (194)$$

Note the plus sign on the left hand side of equations (190) and (193) which indicates that the kinetic energy is positive. For this case  $\beta_{UU} \sim 0$  and  $\beta_{\alpha\alpha} \sim \pi/2$ .

For a slow quantum process in coherent spacetime equation (82) is valid and the momentum equation (188) becomes

$$\bar{p}_{\alpha}^{sc} = i\hbar/\bar{\alpha} \partial U/\partial \theta_{\alpha} \quad (195)$$

which agrees with equations (34) and (143). In the case of a slow quantum process in coherent spacetime equation (82) is valid and Schrödinger's equation (189) becomes

$$-\sum_{\alpha} \sec \beta_{w\alpha w\alpha}^{sc} \alpha^{-2} \partial^2 U/\partial \theta_{\alpha}^2 \exp(j\theta_{\eta\alpha}^{sc}) = \bar{k}^2 U \quad (196)$$

where from equation (90) with  $\theta_U = 0$

$$\theta_{\eta\alpha}^{sc} = \beta_{w\alpha w\alpha}^{sc} - 2\theta_{\alpha} - \pi \quad (197)$$

The small positive angle  $\beta_{w\alpha w\alpha}^{sc}$  is given by

$$\tan \beta_{w\alpha w\alpha}^{sc} = w_{\alpha}^{sc} d\theta_{w\alpha}^{sc}/dw_{\alpha}^{sc} = E_{U\alpha}^{sc}/F_{U\alpha}^{sc} \quad (198)$$

$$\sec \beta_{w\alpha w\alpha}^{sc} = (F_{U\alpha}^{sc})^{-1} [(E_{U\alpha}^{sc})^2 + (F_{U\alpha}^{sc})^2]^{1/2} \quad (199)$$

where

$$E_{U\alpha}^{sc} = -\partial U/\partial \theta_{\alpha} \quad E_{U\alpha}^{sc} \geq 0 \quad (200)$$

$$F_{U\alpha}^{sc} = \partial^2 U/\partial \theta_{\alpha}^2 \quad F_{U\alpha}^{sc} \geq 0 \quad (201)$$

Using equations (198) through (201) allows Schrödinger's equation (196) to be written as

$$-\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{sc})^2 + (F_{U\alpha}^{sc})^2]^{1/2} \exp(j\theta_{\eta\alpha}^{sc}) = \bar{k}^2 U \quad (202)$$

Using equations (98) and (168) shows that equation (202) is equivalent to the following complex number form of Schrödinger's equation

$$\sum_{\alpha} 1/\bar{\alpha}^2 (F_{U\alpha}^{sc} + jE_{U\alpha}^{sc}) = \bar{k}^2 U \quad (203)$$

where in an analogous manner to equation (98) and (169)

$$\bar{\eta}_{\alpha}^{sc} = -1/\bar{\alpha}^2 (F_{U\alpha}^{sc} + jE_{U\alpha}^{sc}) \quad (204)$$

and equation (203) therefore follows from equations (168) and (204).

The alternative description of the second spatial derivative given by equations (87), (88), (96) and (97) allows Schrödinger's equation (202) to be written as

$$\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{sc})^2 + (F_{U\alpha}^{sc})^2]^{1/2} \exp(j\theta_{\eta\alpha}^{sc'}) = \bar{k}^2 U \quad (205)$$

where

$$\begin{aligned} \theta_{\eta\alpha}^{sc'} &= \theta_{\eta\alpha}^{sc} + \pi \\ &= \beta_{wawa}^{sc} - 2\theta_{\alpha} \end{aligned} \quad (206)$$

By assuming that the internal phase angles of all of the component terms of equation (205) are equal, this equation can be written as two approximate scalar forms of Schrödinger's equation

$$\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{sc})^2 + (F_{U\alpha}^{sc})^2]^{1/2} = k^2 U \quad (207)$$

$$\beta_{wawa}^{sc} - 2\theta_{\alpha} = 2\theta_k \quad (208)$$

Equation (208) is the appropriate limiting form of equation (192). A further approximate representation of Schrödinger's equation (205) is obtained from equations (165) and (166) as

$$\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{sc})^2 + (F_{U\alpha}^{sc})^2]^{1/2} = k_s^2 U \quad (209)$$

$$\theta_{\eta\alpha}^{sc'} = 2\theta_{ks} = \theta_E = \theta_V \quad (210)$$

for  $\alpha = x, y$  and  $z$ . Combining equations (206) and (210) gives

$$\beta_{wawa}^{sc} - 2\theta_{\alpha} = 2\theta_{ks} = \theta_E = \theta_V \quad (211)$$

which agrees with equation (194) for the case of a slow quantum process in coherent space. Equation (209) can be obtained directly from equation (193) for

this special case. Equations (207) through (211) give approximate scalar forms of Schrodinger's equation for a slow process in coherent space.

For the one dimensional case Schrödinger's equations (203), (205) and (206) can be written for a slow process in coherent space as

$$1/\bar{x}^2 (F_{Ux}^{sc} + jE_{Ux}^{sc}) = \bar{k}^2 U \quad (212)$$

$$1/x^2 [(E_{Ux}^{sc})^2 + (F_{Ux}^{sc})^2]^{1/2} \exp(j\theta_{\eta x}^{sc'}) = \bar{k}^2 U \quad (213)$$

$$\theta_{\eta x}^{sc'} = \beta_{wxwx}^{sc} - 2\theta_x \quad (214)$$

where

$$\tan \beta_{wxwx}^{sc} = E_{Ux}^{sc}/F_{Ux}^{sc} \quad (215)$$

Schrödinger's equation (213) can be written exactly as

$$[(E_{Ux}^{sc})^2 + (F_{Ux}^{sc})^2]^{1/2} = k_x^2 U \quad (216)$$

$$\beta_{wxwx}^{sc} - 2\theta_x = 2\theta_k \quad (217)$$

where  $\theta_k$  and  $k$  are given by equations (163) and (164). Approximate forms of Schrödinger's equations (216) and (217) are written as

$$[(E_{Ux}^{sc})^2 + (F_{Ux}^{sc})^2]^{1/2} = k_s^2 U \quad (218)$$

$$\beta_{wxwx}^{sc} - 2\theta_x = 2\theta_{ks} \quad (219)$$

where  $k_s$  and  $\theta_{ks}$  are given by equations (165) and (166) and where  $x = \text{constant}$ . Equations (216) through (219) are the scalar forms of Schrödinger's equation in one dimension for the case of a slow quantum process in coherent space.

Case c. Ultrafast Quantum Processes in Incoherent Space.

The momentum equation for this case is obtained from equations (143), (146) and (149) to be

$$\bar{p}_\alpha \bar{U} = -i\hbar \csc \beta_{UU} \cos \beta_{\alpha\alpha} U \partial \theta_U / \partial \alpha \exp(j\theta_{w\alpha}) \quad (220)$$

or

$$\bar{p}_\alpha = -i\hbar \csc \beta_{UU} \cos \beta_{\alpha\alpha} \partial \theta_U / \partial \alpha \exp(j\theta_{w\alpha}') \quad (221)$$

where

$$\theta_{w\alpha}' = \theta_{w\alpha} - \theta_U \quad (222)$$

$$= \beta_{UU} - \theta_\alpha - \beta_{\alpha\alpha}$$

where  $\bar{U}$  has been divided out of equation (221). The proper form of Schrödinger's equation (167) that is suitable for passing to the limiting case of an ultrafast quantum process in incoherent spacetime can be obtained from equations (49), (59) and (62) as

$$- \sum_{\alpha} \csc \beta_{\alpha\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{UU} \partial \theta_U / \partial \alpha \partial \theta_{\alpha\alpha} / \partial \alpha \exp(j\lambda_{U\alpha}) = \bar{k}^2 \quad (223)$$

where the complex number wave function  $\bar{U}$  has been divided out of equation (223) and where equation (60) gives

$$\partial \theta_{\alpha\alpha} / \partial \alpha = \partial / \partial \alpha (\theta_U + \beta_{UU} - \theta_{\alpha} - \beta_{\alpha\alpha}) \quad (224)$$

and where  $\lambda_{U\alpha}$  is obtained from equation (62) as

$$\begin{aligned} \lambda_{U\alpha} &= \theta_{\alpha\alpha} - \theta_U \\ &= \beta_{UU} + \beta_{\alpha\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \end{aligned} \quad (225)$$

Schrödinger's equation (223) can be represented by two approximate scalar equations by using the representation in equations (64) and (65) and assuming the equality of the internal phase angles of each term in equation (223) with the result that Schrödinger's equation becomes

$$\sum_{\alpha} \csc \beta_{\alpha\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{UU} \partial \theta_U / \partial \alpha \partial \theta_{\alpha\alpha} / \partial \alpha = k^2 \quad (226)$$

$$\beta_{UU} + \beta_{\alpha\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) - \pi = 2\theta_k \quad (227)$$

where  $\theta_k$  and  $k$  are given by equations (163) and (164). A further approximation gives Schrödinger's equation as

$$\sum_{\alpha} \csc \beta_{\alpha\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{UU} \partial \theta_U / \partial \alpha \partial \theta_{\alpha\alpha} / \partial \alpha = k_s^2 \quad (228)$$

$$\beta_{UU} + \beta_{\alpha\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) - \pi = 2\theta_{ks} = \theta_V = \theta_E \quad (229)$$

where  $k_s$  and  $\theta_{ks}$  are given by equations (165) and (166).

For an ultrafast quantum process in incoherent space equation (57) is valid and the momentum equation (221) becomes

$$\bar{p}_{\alpha}^{ui} = -ij\hbar \partial \theta_U / \partial \alpha \quad (230)$$

where for this case  $\beta_{UU} = \pi/2$ ,  $\theta_{\alpha} = 0$ ,  $\beta_{\alpha\alpha} = 0$  and  $\theta_{\alpha\alpha} = \pi/2$ , and where  $\bar{U}$  has been divided out of equation (220), and which agrees with equations (31) and (143). In the case of an ultrafast process in incoherent spacetime the combination of equations (57), (223) and (225) give Schrödinger's equation as

$$- \sum_{\alpha} \csc \beta_{\alpha\alpha}^{ui} (\partial \theta_U / \partial \alpha)^2 \exp(j\lambda_{U\alpha}^{ui}) = \bar{k}^2 \quad (231)$$

$$\lambda_{U\alpha}^{ui} = \theta_{\eta\alpha}^{ui} - \theta_U \quad (232)$$

$$= \pi/2 + \beta_{waw\alpha}^{ui} = \pi + \kappa_{U\alpha} \quad (233)$$

where  $\beta_{waw\alpha}^{ui}$  is obtained from equation (71) as

$$\tan \beta_{waw\alpha}^{ui} = E_{U\alpha}^{ui}/F_{U\alpha}^{ui} \quad (234)$$

$$\csc \beta_{waw\alpha}^{ui} = (E_{U\alpha}^{ui})^{-1} [(E_{U\alpha}^{ui})^2 + (F_{U\alpha}^{ui})^2]^{1/2} \quad (235)$$

$$E_{U\alpha}^{ui} = (\partial\theta_U/\partial\alpha)^2 \quad E_{U\alpha}^{ui} \geq 0 \quad (236)$$

$$F_{U\alpha}^{ui} = \partial^2\theta_U/\partial\alpha^2 \quad F_{U\alpha}^{ui} \leq 0 \quad (237)$$

Then equation (233) gives

$$\begin{aligned} \kappa_{U\alpha} &= \tan^{-1}(|F_{U\alpha}^{ui}|/E_{U\alpha}^{ui}) \\ &= \beta_{waw\alpha}^{ui} - \pi/2 \end{aligned} \quad (238)$$

and from equation (78) it follows that

$$\theta_{\eta\alpha}^{ui} = \theta_U + \kappa_{U\alpha} + \pi \quad (239)$$

while equation (81) gives

$$\bar{\eta}_{\alpha}^{ui} = -\bar{U}(E_{U\alpha}^{ui} - jF_{U\alpha}^{ui}) \quad (240)$$

which is a special case of equation (169).

Combining equations (231), (235) and (236) gives the Schrödinger equation as

$$-\sum_{\alpha} [(E_{U\alpha}^{ui})^2 + (F_{U\alpha}^{ui})^2]^{1/2} \exp(j\lambda_{U\alpha}^{ui}) = \bar{k}^2 \quad (241)$$

or equivalently using equations (232) and (233)

$$\sum_{\alpha} [(E_{U\alpha}^{ui})^2 + (F_{U\alpha}^{ui})^2]^{1/2} \exp(j\kappa_{U\alpha}) = \bar{k}^2 \quad (242)$$

Using equations (168) and (240) gives Schrödinger's equation for this case also as

$$\sum_{\alpha} (E_{U\alpha}^{ui} - jF_{U\alpha}^{ui}) = \bar{k}^2 \quad (243)$$

Schrödinger's equation (242) can be written in an approximate form by assuming that each of the phase angles in the sum on the left hand side of equation (242) are equal, with the result that

$$\sum_{\alpha} [(E_{U\alpha}^{ui})^2 + (F_{U\alpha}^{ui})^2]^{1/2} = k^2 \quad (244)$$

$$\kappa_{U\alpha} = 2\theta_k \quad (245)$$

where  $\kappa_{U\alpha}$  is given by equation (238). A further approximate form of Schrödinger's equation for this case is obtained from equations (165) and (166) as

$$\sum_{\alpha} [(E_{U\alpha}^{ui})^2 + (F_{U\alpha}^{ui})^2]^{1/2} = k_s^2 \quad (246)$$

$$\kappa_{U\alpha} = 2\theta_{ks} = \theta_V = \theta_E \quad (247)$$

where  $\kappa_{U\alpha}$  is a small angle. Equations (244) and (245) agree with equations (226) and (227), and equations (246) and (247) agree with equations (228) and (229) in the limiting case of an ultrafast process in coherent space.

For the one dimensional case Schrödinger's equations (242) and (243) for an ultrafast process in incoherent space are written as

$$[(E_{Ux}^{ui})^2 + (F_{Ux}^{ui})^2]^{1/2} \exp(j\kappa_{Ux}) = \bar{k}^2 \quad (248)$$

$$E_{Ux}^{ui} - jF_{Ux}^{ui} = \bar{k}^2 \quad (249)$$

Schrödinger's equations (248) or (249) can also be written as

$$[(E_{Ux}^{ui})^2 + (F_{Ux}^{ui})^2]^{1/2} = k^2 \quad (250)$$

$$\kappa_{Ux} = 2\theta_k \quad (251)$$

where

$$\tan \kappa_{Ux} = |F_{Ux}^{ui}|/E_{Ux}^{ui} \quad (252)$$

and  $\theta_k$  and  $k$  are given by equations (163) and (164). The approximate forms of Schrödinger's equations (250) and (251) are written as

$$[(E_{Ux}^{ui})^2 + (F_{Ux}^{ui})^2]^{1/2} = k_s^2 \quad (253)$$

$$\kappa_{Ux} = 2\theta_{ks} \quad (254)$$

where  $k_s$  and  $\theta_{ks}$  are given by equations (165) and (166).

Case d. Ultrafast Quantum Processes in Coherent Spacetime.

The quantum mechanical expression for the momentum of a particle that is suitable for this case is obtained from equations (143), (148) and (149) as

$$\bar{p}_\alpha \bar{U} = -i\hbar \csc \beta_{UU} \sin \beta_{\alpha\alpha} U/\alpha \partial \theta_U / \partial \theta_\alpha \exp(j\theta_{w\alpha}) \quad (255)$$

or equivalently

$$\bar{p}_\alpha = -i\hbar \csc \beta_{UU} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial \theta_U / \partial \theta_\alpha \exp(j\theta'_{w\alpha}) \quad (256)$$

where

$$\theta'_{w\alpha} = \theta_{w\alpha} - \theta_U = \beta_{UU} - \beta_{\alpha\alpha} - \theta_\alpha \quad (257)$$

and where  $\bar{U}$  has been divided out of equation (255). The form of Schrödinger's equation (167) that is required for the passage to the limit of an ultrafast quantum process in coherent spacetime can be obtained from equations (101), (102) and (225) as

$$-\sum_{\alpha} \csc \beta_{w\alpha w\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{UU} \alpha^{-2} \partial \theta_U / \partial \theta_\alpha \partial \theta_{w\alpha} / \partial \theta_\alpha \exp(j\lambda_{U\alpha}) = \bar{k}^2 \quad (258)$$

where the complex number wave function  $\bar{U}$  has been divided out of equation (258),  $\lambda_{U\alpha}$  is given by equation (225), and where equation (149) gives

$$\partial \theta_{w\alpha} / \partial \theta_\alpha = \partial \theta_U / \partial \theta_\alpha - 1 + \partial / \partial \theta_\alpha (\beta_{UU} - \beta_{\alpha\alpha}) \quad (259)$$

Schrödinger's equation (258) can be written as two approximate scalar equations by using equations (115) and (116) and by assuming that the phase angles of each term in the sum of equation (258) are equal with the result that Schrödinger's equation becomes

$$\sum_{\alpha} \csc \beta_{w\alpha w\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{UU} \alpha^{-2} \partial \theta_U / \partial \theta_\alpha \partial \theta_{w\alpha} / \partial \theta_\alpha = k^2 \quad (260)$$

$$\beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_\alpha + \beta_{\alpha\alpha}) + \pi = 2\theta_k \quad (261)$$

A further approximation for Schrödinger's equation follows by using equations (165) and (166) which gives

$$\sum_{\alpha} \csc \beta_{w\alpha w\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{UU} \alpha^{-2} \partial \theta_U / \partial \theta_\alpha \partial \theta_{w\alpha} / \partial \theta_\alpha = k_s^2 \quad (262)$$

$$\beta_{UU} + \beta_{w\alpha w\alpha} - 2(\theta_\alpha + \beta_{\alpha\alpha}) + \pi = 2\theta_{ks} \quad (263)$$

The kinetic energy term of Schrödinger's equation in (260) and (262) appears with a positive sign as should be the case.

For the case of an ultrafast process in coherent space equation (99) is valid and the momentum equation (256) becomes



$$\begin{aligned}\bar{p}_\alpha^{uc} &= -i\hbar\alpha^{-1}\partial\theta_U/\partial\theta_\alpha \exp(-j\theta_\alpha) \\ &= -i\hbar/\bar{\alpha} \partial\theta_U/\partial\theta_\alpha\end{aligned}\quad (264)$$

Equation (264) also follows directly from equation (152) for the special case of an ultrafast process. For the case of an ultrafast quantum process in coherent spacetime, equations (99), (258) and (259) gives Schrödinger's equation as

$$-\int_{\bar{\alpha}} \csc \beta_{wawa}^{uc} \alpha^{-2} \partial\theta_U/\partial\theta_\alpha (\partial\theta_U/\partial\theta_\alpha - 1) \exp(j\lambda_{U\alpha}^{uc}) = \bar{k}^2 \quad (265)$$

where equations (99), (128), (130), (133) and (225) give

$$\lambda_{U\alpha}^{uc} = \theta_{\eta\alpha}^{uc} - \theta_U \quad (266)$$

$$= \theta_{\eta\alpha}^{uc'} - \theta_U - \pi \quad (267)$$

$$= \beta_{wawa}^{uc} - 2\theta_\alpha - \pi/2 \quad (268)$$

$$= \delta_{U\alpha} - 2\theta_\alpha - \pi \quad (269)$$

where  $\theta_{\eta\alpha}^{uc}$  and  $\theta_{\eta\alpha}^{uc'}$  are given by equations (130) through (133). Equations (125) through (129) give

$$\tan \beta_{wawa}^{uc} = E_{U\alpha}^{uc}/F_{U\alpha}^{uc} \quad (270)$$

$$\csc \beta_{wawa}^{uc} = (E_{U\alpha}^{uc})^{-1} [(E_{U\alpha}^{uc})^2 + (F_{U\alpha}^{uc})^2]^{1/2} \quad (271)$$

$$\beta_{wawa}^{uc} = -\pi/2 + \delta_{U\alpha} \quad (272)$$

$$\tan \delta_{U\alpha} = F_{U\alpha}^{uc}/|E_{U\alpha}^{uc}| \quad (273)$$

where from equations (121) and (122)

$$E_{U\alpha}^{uc} = \partial\theta_U/\partial\theta_\alpha (\partial\theta_U/\partial\theta_\alpha - 1) \quad E_{U\alpha}^{uc} \leq 0 \quad (274)$$

$$F_{U\alpha}^{uc} = \partial^2\theta_U/\partial\theta_\alpha^2 \quad F_{U\alpha}^{uc} \geq 0 \quad (275)$$

where  $\delta_{U\alpha}$  is a small number. It is also useful to define the following quantity

$$\lambda_{U\alpha}^{uc'} = \lambda_{U\alpha}^{uc} + \pi \quad (276)$$

$$= \theta_{\eta\alpha}^{uc'} - \theta_U \quad (277)$$

$$= \delta_{U\alpha} - 2\theta_\alpha \quad (278)$$

where  $\lambda_{U\alpha}^{uc'}$  is a small angle.

Combining equations (121), (126) and (265) gives Schrödinger's equation as

$$-\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{uc})^2 + (F_{U\alpha}^{uc})^2]^{1/2} \exp(j\lambda_{U\alpha}^{uc}) = \bar{k}^2 \quad (279)$$

Combining equations (276) and (279) give Schrödinger's equation as

$$\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{uc})^2 + (F_{U\alpha}^{uc})^2]^{1/2} \exp(j\lambda_{U\alpha}^{uc'}) = \bar{k}^2 \quad (280)$$

For the case of an ultrafast process in coherent space it is easy to show that equations (153) through (155) give

$$\bar{p}_{\alpha}^2 = \hbar^2 / \bar{\alpha}^2 (-E_{U\alpha}^{uc} + jF_{U\alpha}^{uc}) \quad (281)$$

so that Schrödinger's equation (280) is equivalent to

$$\sum_{\alpha} 1/\bar{\alpha}^2 (-E_{U\alpha}^{uc} + jF_{U\alpha}^{uc}) = \bar{k}^2 \quad (282)$$

which is a result that can also be obtained from equation (168) by noting that in an analogous fashion to equation (135)

$$\bar{\eta}_{\alpha}^{uc} = -\bar{U}/\bar{\alpha}^2 (-E_{U\alpha}^{uc} + jF_{U\alpha}^{uc}) \quad (283)$$

Schrödinger's equation (282) can be rewritten as

$$\sum_{\alpha} \alpha^{-2} \exp(-2j\theta_{\alpha}) [\partial\theta_U / \partial\theta_{\alpha} (1 - \partial\theta_U / \partial\theta_{\alpha}) + j\partial^2\theta_U / \partial\theta_{\alpha}^2] = \bar{k}^2 \quad (284)$$

An approximate set of scalar Schrödinger equations that are equivalent to equations (280) and (282) are written as

$$\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{uc})^2 + (F_{U\alpha}^{uc})^2]^{1/2} = k^2 \quad (285)$$

$$\lambda_{U\alpha}^{uc'} = 2\theta_k \quad (286)$$

Combining equations (278) and (286) gives

$$\delta_{U\alpha} - 2\theta_{\alpha} = 2\theta_k \quad (287)$$

Equations (285) and (287) agree with the limiting forms of Schrödinger's equations (260) and (261) for the case of an ultrafast process in coherent space. A further approximate form of Schrödinger's equation follows from equations (165), (166), (285) and (286) as

$$\sum_{\alpha} \alpha^{-2} [(E_{U\alpha}^{uc})^2 + (F_{U\alpha}^{uc})^2]^{1/2} = k_s^2 \quad (288)$$

$$\delta_{U\alpha} - 2\theta_{\alpha} = 2\theta_{ks} = \theta_V = \theta_E \quad (289)$$

For one space dimension Schrödinger's equation for an ultrafast process in coherent space is given by any of the following three forms

$$1/\bar{x}^2 [d\theta_U/d\theta_x (1 - d\theta_U/d\theta_x) + j d^2\theta_U/d\theta_x^2] = \bar{k}^2 \quad (290)$$

$$x^{-2} \exp(-2j\theta_x) [d\theta_U/d\theta_x (1 - d\theta_U/d\theta_x) + j d^2\theta_U/d\theta_x^2] = \bar{k}^2 \quad (291)$$

$$1/\bar{x}^2 (-E_{Ux}^{uc} + jF_{Ux}^{uc}) = \bar{k}^2 \quad (292)$$

where  $x = \text{constant}$ . Schrödinger's equation (292) can also be written as

$$[(E_{Ux}^{uc})^2 + (F_{Ux}^{uc})^2]^{1/2} = k_s^2 x^2 \quad (293)$$

$$\delta_{Ux} - 2\theta_x = 2\theta_k \quad (294)$$

where

$$\tan \delta_{Ux} = F_{Ux}^{uc} / |E_{Ux}^{uc}| \quad (295)$$

The approximate form of Schrödinger's equations (293) and (294) are written as

$$[(E_{Ux}^{uc})^2 + (F_{Ux}^{uc})^2]^{1/2} = k_s^2 x^2 \quad (296)$$

$$\delta_{Ux} - 2\theta_x = 2\theta_{ks} = \theta_V = \theta_E \quad (297)$$

In general  $k = k(x, \theta_x)$  while  $k_s = k_s(x) = \text{constant}$ .

The kinetic energy term for the case of an ultrafast quantum process in coherent space can be obtained as a special case of the more general situation of an arbitrary quantum process in coherent space which is represented by equation (155) of this paper or equation (250) of Reference 26, so that Schrödinger's equation for an arbitrary process in coherent space is given by

$$\sum_{\alpha} 1/\bar{\alpha}^2 (\partial^2 \bar{U} / \partial \theta_{\alpha}^2 - j \partial \bar{U} / \partial \theta_{\alpha}) = \bar{k}^2 \bar{U} \quad (298)$$

Schrödinger's equation (298) can be specialized to the case of an ultrafast process in coherent space by noting that for this case

$$\partial \bar{U} / \partial \theta_{\alpha} = j \bar{U} \partial \theta_U / \partial \theta_{\alpha} \quad (299)$$

$$\partial^2 \bar{U} / \partial \theta_{\alpha}^2 = \bar{U} [ - (\partial \theta_U / \partial \theta_{\alpha})^2 + j \partial^2 \theta_U / \partial \theta_{\alpha}^2 ] \quad (300)$$

Combining equations (298) through (300) immediately gives the Schrödinger's equation (282).

### C. Examples of Complex Number Potential Functions.

This section considers several complex number potential functions which

are of interest to the case of ultrafast processes in coherent space. A simple solution to Schrödinger's equation (290) or (291) can be obtained for the special simplifying condition

$$\bar{x}^2 k^2 = x^2 k^2 = a^2 = \text{real constant} \quad (301)$$

which may be approximately true over a limited range of space. Equation (301) gives the constant wave number as

$$k = a/x \quad (302)$$

where  $x = \text{constant}$  for coherent space. Equation (157) shows that the conditions in equations (301) and (302) are exactly valid for a choice of the complex number potential as

$$\bar{V} = \bar{E} - [a^2 \hbar^2 / (2\mu)] / \bar{x}^2 \quad (303)$$

Then Schrödinger's equation (290) is written as

$$d\theta_U / d\theta_x (1 - d\theta_U / d\theta_x) + j d^2 \theta_U / d\theta_x^2 = a^2 \quad (304)$$

which gives

$$d\theta_U / d\theta_x (1 - d\theta_U / d\theta_x) = -E_{Ux}^{uc} = a^2 \quad (305)$$

$$d^2 \theta_U / d\theta_x^2 = F_{Ux}^{uc} = 0 \quad (306)$$

Equation (306) suggests a solution of the form

$$\theta_U = f\theta_x + g \quad (307)$$

where  $f$  and  $g$  are constants. Combining equations (305) and (307) gives

$$f(1 - f) = a^2 = k^2 x^2 \quad (308)$$

whose solution is

$$\begin{aligned} f &= 1/2 \pm 1/2(1 - 4a^2)^{1/2} \\ &= 1/2 \pm 1/2(1 - 4k^2 x^2)^{1/2} \end{aligned} \quad (309)$$

where  $x = \text{constant}$

One example of the complex number potential for the case of an internal phase harmonic oscillator in one space dimension is given by

$$\begin{aligned} \bar{V} &= 1/2 \bar{K} \bar{x}^2 & \bar{K} &= K \exp(j\theta_K) \\ &= 1/2 K x^2 \exp[j(\theta_K + 2\theta_x)] \end{aligned} \quad (310)$$

where  $\bar{K}$  = constant and  $x$  = constant so that  $\theta_x$  is the dynamical variable. The scalar equations corresponding to equation (310) are

$$V = 1/2Kx^2 = \text{constant} \quad (311)$$

$$\theta_V = \theta_K + 2\theta_x \quad (312)$$

The complex number wave number  $\bar{k}$  for the simple harmonic oscillator is obtained from equations (157) and (310) to be

$$\bar{k}^2 = 2\mu/\hbar^2(\bar{E} - 1/2\bar{K}x^2) \quad (313)$$

while equations (159) through (164) determine  $k(\theta_x)$  and  $\theta_k(\theta_x)$ . Schrödinger's equation (290) becomes for this case

$$1/\bar{x}^2[d\theta_U/d\theta_x(1 - d\theta_U/d\theta_x) + jd^2\theta_U/d\theta_x^2] = 2\mu/\hbar^2(\bar{E} - 1/2\bar{K}x^2) \quad (314)$$

whose solution is not simply obtained.

Another possible type of simple harmonic oscillator potential that may describe internal space motions is given by

$$\bar{V}_i = 1/2\bar{K}_{ix}\theta_x^2 = V_i \exp(j\theta_{Vi}) \quad (315)$$

where

$$\bar{K}_{ix} = K_{ix} \exp(j\theta_{Kix}) \quad (316)$$

so that

$$V_i = 1/2K_{ix}\theta_x^2 \quad \theta_{Vi} = \theta_{Kix} \quad (317)$$

Then equation (157) gives the wave number as

$$\bar{k}^2 = 2\mu/\hbar^2(\bar{E}_i - 1/2\bar{K}_i\theta_x^2) \quad (318)$$

from which  $k(\theta_x)$  and  $\theta_k(\theta_x)$  can be calculated by equations (159) through (164). The Schrödinger equation (290) for this case is written as

$$1/\bar{x}^2[d\theta_U/d\theta_x(1 - d\theta_U/d\theta_x) + jd^2\theta_U/d\theta_x^2] = 2\mu/\hbar^2(\bar{E}_i - 1/2\bar{K}_i\theta_x^2) \quad (319)$$

Equation (319) is equivalent to equation (372) of Reference 27 for the special case of a coherent wave function that is associated with an ultrafast quantum process.

**3. TIME DEPENDENT SCHRÖDINGER EQUATION IN BROKEN SYMMETRY CARTESIAN SPACETIME.** This section considers the time dependent Schrödinger equation for slow and ultrafast quantum processes in spacetime with broken internal symmetries. Only cartesian coordinates are considered. The complex number time

dependent Schrödinger equation is written as the following generalization of the standard equation<sup>22-26</sup>

$$- \hbar^2 / (2\mu) \bar{\nabla}^2 \bar{\Psi} + \bar{V} \bar{\Psi} = i \hbar \partial \bar{\Psi} / \partial \bar{t} \quad (320)$$

where the wave function and potential are written as

$$\bar{\Psi} = \Psi \exp(j\theta_{\Psi}) \quad (321)$$

$$\bar{V} = V \exp(j\theta_V) \quad (322)$$

For cartesian coordinates equation (320) becomes

$$- \hbar^2 / (2\mu) \sum_{\alpha} \partial^2 \bar{\Psi} / \partial \bar{\alpha}^2 + \bar{V} \bar{\Psi} = i \hbar \partial \bar{\Psi} / \partial \bar{t} \quad (323)$$

where  $\bar{\Psi} = \bar{\Psi}(\bar{\alpha}, \bar{t})$ . Combining equations (16), (49) and (323) gives

$$- \hbar^2 / (2\mu) \sum_{\alpha} \bar{\xi}_{\alpha} + \bar{V} \bar{\Psi} = i \hbar \bar{u} \quad (324)$$

where  $\bar{u}$  and  $\bar{\xi}_{\alpha}$  are given by equations (16) and (49). Equation (324) can be rewritten as

$$- \hbar^2 / (2\mu) \sum_{\alpha} \xi_{\alpha} \exp(j\theta_{\xi_{\alpha}}) + V \Psi \exp[j(\theta_V + \theta_{\Psi})] = i \hbar u \exp(j\theta_u) \quad (325)$$

where  $\xi_{\alpha}$  and  $u$  are evaluated in Section 1 for various spacetime cases. In one dimension equation (325) becomes

$$- \hbar^2 / (2\mu) \xi_x \exp(j\theta_{\xi_x}) + V \Psi \exp[j(\theta_V + \theta_{\Psi})] = i \hbar u \exp(j\theta_u) \quad (326)$$

where  $u$  and  $\xi_x$  must be evaluated according to the relevant spacetime conditions as in Section 1.

The exact solution of equation (325) requires that the real and imaginary parts of each component term be taken, but this leads to very complicated equations. A simpler but approximate procedure of solving equation (325) assumes that the internal phase angles of each term in equation (325) are equal. Then because the internal phase angles of the second derivative can have a  $(0, \pm\pi)$  term included as described in Section 1 in equations (54), (65), (88) and (116), and the internal phase angles of the first derivative with respect to time can have a term  $(0, \pm\pi/2)$  added as described by equations (39), (41), (44) and (47), the solution of equation (325) requires that eight possible cases be considered individually: slow and fast processes, coherent and incoherent space, and coherent and incoherent time.<sup>21</sup> Therefore equation (325) can be written analogously to equations (170) and (171) as the following approximations that are obtained by assuming that the internal phase angles of each component term are equal<sup>21</sup>

$$- (\pm 1) \hbar^2 / (2\mu) \sum_{\alpha} \xi_{\alpha} + V\Psi = i\hbar u \quad (327)$$

$$\theta_{\xi\alpha} + (0, \pm\pi) = \theta_V + \theta_{\Psi} = \theta_u + (0, \pm\pi/2) \quad (328)$$

From equations (26) and (54) it follows that equation (328) can be written in analogy to equation (174) as the following approximation

$$\beta_{\Psi\Psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + (0, \pm\pi) = \theta_V = \beta_{\Psi\Psi} - \theta_t^{\Psi} - \beta_{tt}^{\Psi} + (0, \pm\pi/2) \quad (329)$$

A comparison of equations (174) and (329) shows that the following approximations are valid for a stationary state

$$\beta_{\Psi\Psi} - \theta_t^{\Psi} - \beta_{tt}^{\Psi} + (0, \pm\pi/2) = 2\theta_k \sim 2\theta_{ks} = \theta_V = \theta_E \quad (330)$$

and

$$\beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + (0, \pm\pi) = -\theta_t^{\Psi} - \beta_{tt}^{\Psi} + (0, \pm\pi/2) \quad (331)$$

The time dependent Schrödinger equation given by equations (327) through (331) will now be delineated for the eight possible states that are associated with a slow or ultrafast quantum process, coherent or incoherent spatial coordinate variation, and coherent or incoherent time variation.<sup>27</sup> Reference 27 on wave propagation establishes the basis of development for this section.

#### Case a. Slow Quantum Process, Incoherent Space and Incoherent Time.

This section develops a general form of the time dependent Schrödinger equation that can be used to obtain the limiting case of a slow quantum process that occurs in **incoherent** space and **incoherent** time which is described by

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \theta_t^{\Psi} = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (332)$$

For the case when all internal phase angles are small, equations (327) and (328) are written as

$$- \hbar^2 / (2\mu) \sum_{\alpha} \xi_{\alpha} + V\Psi = i\hbar u \quad (333)$$

$$\theta_{\xi\alpha} = \theta_V + \theta_{\Psi} = \theta_u \quad (334)$$

For this case equations (22), (52) and (333) give

$$\begin{aligned} & - \hbar^2 / (2\mu) \sum_{\alpha} \sec \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} \partial / \partial \alpha (\sec \beta_{\Psi\Psi} \cos \beta_{\alpha\alpha} \partial \Psi / \partial \alpha) \\ & + V\Psi = i\hbar \sec \beta_{\Psi\Psi} \cos \beta_{tt}^{\Psi} \partial \Psi / \partial t \end{aligned} \quad (335)$$

where  $\beta_{v\alpha v\alpha}$  is given by equations (17), (21) and (55). Combining equations (329) and (334) gives

$$\beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \theta_v = \beta_{\psi\psi} - \theta_t^{\psi} - \beta_{tt}^{\psi} \quad (336)$$

For the case of a slow quantum process in incoherent space and incoherent time equation (332) is valid and equation (335) becomes the standard time dependent Schrödinger equation

$$-\hbar^2/(2\mu) \sum_{\alpha} \partial^2 \psi / \partial \alpha^2 + V\psi = i\hbar \partial \psi / \partial t \quad (337)$$

while equation (336) gives a null result. A stationary state follows from the choice

$$\psi = U(\alpha) \exp(-iEt/\hbar) \quad (338)$$

which combined with equation (337) yields the stationary state Schrödinger equation (187).

#### Case b. Slow Quantum Process, Coherent Space and Incoherent Time.

A general form of the time dependent Schrödinger equation is developed that can be used to deduce the limiting form of this equation for a slow quantum process in coherent space and incoherent time which is characterized by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad \theta_t^{\psi} = 0 \quad \beta_{tt}^{\psi} = 0 \quad (339)$$

For the general case equations (22) and (84) can be combined with equation (325) to give the following time dependent Schrödinger equation

$$\begin{aligned} & -\hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial \theta_{\alpha} (\sec \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial \psi / \partial \theta_{\alpha}) \exp(j\theta_{\xi\alpha}) \\ & + \bar{V}\bar{\psi} = i\hbar \sec \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \partial \psi / \partial t \exp(j\theta_u) \end{aligned} \quad (340)$$

An approximate representation of equations (323) or (340) is given by

$$\hbar^2/(2\mu) \sum_{\alpha} \xi_{\alpha} + V\psi = i\hbar u \quad (341)$$

$$\theta_{\xi\alpha} + \pi = \theta_v + \theta_{\psi} = \theta_u \quad (342)$$

or equivalently

$$\begin{aligned} & \hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial \theta_{\alpha} (\sec \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial \psi / \partial \theta_{\alpha}) \\ & + V\psi = i\hbar \sec \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \partial \psi / \partial t \end{aligned} \quad (343)$$

$$\beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = \theta_v = \beta_{\psi\psi} - \theta_t^{\psi} - \beta_{tt}^{\psi} \quad (344)$$

which allows the kinetic energy term to appear as a positive quantity.



The time dependent Schrödinger equation which is the limiting form of equation (340) for the case of a slow quantum process in coherent space and incoherent time follows from equation (98), (136) and (339) as

$$\hbar^2/(2\mu) \sum_{\alpha} 1/\alpha^2 (F_{\Psi\alpha}^{sc} + jE_{\Psi\alpha}^{sc}) + \bar{V}\Psi = i\hbar\partial\Psi/\partial t \quad (345)$$

For this special case of a slow quantum process in coherent space and incoherent time equations (339), (343) and (344) give the following approximate time dependent Schrödinger equations

$$\hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{\Psi\alpha}^{sc} \alpha^{-2} \partial^2\Psi/\partial\theta_{\alpha}^2 + V\Psi = i\hbar\partial\Psi/\partial t \quad (346)$$

$$\beta_{\Psi\alpha}^{sc} - 2\theta_{\alpha} = \theta_V = \theta_E = 0 \quad (347)$$

where  $\beta_{\Psi\alpha}^{sc}$  is a small positive angle given by equations (91) through (94). Equation (346) can also be written as

$$\hbar^2/(2\mu) \sum_{\alpha} \alpha^{-2} [(E_{\Psi\alpha}^{sc})^2 + (F_{\Psi\alpha}^{sc})^2]^{1/2} + V\Psi = i\hbar\partial\Psi/\partial t \quad (348)$$

The approximate Schrödinger equations (348) and (347) also follow directly from equation (345). For a stationary state the wave function has the form

$$\Psi = U \exp(-iEt/\hbar) \quad (349)$$

and when placed in the time dependent Schrödinger equation (345) yields the stationary state equation (203) while equation (348) becomes equation (209).

#### Case c. Slow Quantum Process, Incoherent Space and Coherent Time.

This section derives the proper form of the time dependent Schrödinger equation for passing to the limit of a slow quantum process occurring in incoherent space and coherent time which is described by

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \beta_{tt}^{\Psi} = \pi/2 \quad (350)$$

Equations (24) and (52) are combined with equation (325) to give the following Schrödinger equation

$$\begin{aligned} & - \hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{\Psi\alpha} \cos \beta_{\alpha\alpha} \partial/\partial\alpha (\sec \beta_{\Psi\Psi} \cos \beta_{\alpha\alpha} \partial\Psi/\partial\alpha) \exp(j\theta_{\xi\alpha}) \\ & + \bar{V}\Psi = i\hbar \sec \beta_{\Psi\Psi} \sin \beta_{tt}^{\Psi} t^{-1} \partial\Psi/\partial\theta_t^{\Psi} \exp(j\theta_u) \end{aligned} \quad (351)$$

An approximate representation of the Schrödinger equation (351) is given by

$$- \hbar^2/(2\mu) \sum_{\alpha} \xi_{\alpha} + V\Psi = i\hbar u \quad (352)$$

$$\theta_{\xi\alpha} = \theta_V + \theta_{\Psi} = \theta_u + \pi/2 \quad (353)$$

or equivalently using equations (24), (26), (52) and (54)

$$- \hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{\text{vava}} \cos \beta_{\alpha\alpha} \partial/\partial\alpha (\sec \beta_{\psi\psi} \cos \beta_{\alpha\alpha} \partial\psi/\partial\alpha) \quad (354)$$

$$+ V\psi = i\hbar \sec \beta_{\psi\psi} \sin \beta_{\text{tt}}^{\psi} t^{-1} \partial\psi/\partial\theta_t^{\psi}$$

$$\beta_{\psi\psi} + \beta_{\text{vava}} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \theta_V = \beta_{\psi\psi} - \theta_t^{\psi} - \beta_{\text{tt}}^{\psi} + \pi/2 \quad (355)$$

where the right hand sides of equations (352) and (354) are real numbers, and where  $\theta_{\psi}$  has been subtracted from both sides of equation (355). The  $\pi/2$  term in equations (353) and (355) is required to pass to the limit of a slow quantum process in incoherent space and coherent time.

For the special case of a slow quantum process in incoherent space and coherent time equations (350), (354) and (355) give Schrödinger's equation as

$$- \hbar^2/(2\mu) \sum_{\alpha} \partial^2\psi/\partial\alpha^2 + V\psi = i\hbar t^{-1} \partial\psi/\partial\theta_t^{\psi} \quad (356)$$

The phase angle equation (355) becomes in the limit

$$0 \sim \theta_V \sim \theta_E = - \theta_t^{\psi} \quad (357)$$

Consider now the wave function for the stationary state for the slow process in incoherent space and coherent time which is taken to have the following form

$$\psi = U \exp(-iEt\theta_t^{\psi}/\hbar) \quad (358)$$

where  $t = \text{constant}$ . Combining equations (356) and (358) gives the Schrödinger equation for the stationary state as

$$- \hbar^2/(2\mu) \sum_{\alpha} \partial^2 U/\partial\alpha^2 + VU = EU \quad (359)$$

which is the standard time independent Schrödinger equation.

Case d. Slow Quantum Process, Coherent Space and Coherent Time.

This section obtains the required form of the time dependent Schrödinger equation that is needed to obtain the limiting condition of a slow quantum process in coherent space and coherent time which is specified by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad \beta_{\text{tt}}^{\psi} = \pi/2 \quad (360)$$

Equations (24), (84) and (325) gives Schrödinger's equation as

$$- \hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{\text{vava}} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\sec \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial\psi/\partial\theta_{\alpha}) \exp(j\theta_{\xi\alpha}) \quad (361)$$

$$+ \bar{V}\bar{\psi} = i\hbar \sec \beta_{\psi\psi} \sin \beta_{\text{tt}}^{\psi} t^{-1} \partial\psi/\partial\theta_t^{\psi} \exp(j\theta_u)$$

An approximate representation of Schrödinger's equation (361) is given by

$$\hbar^2/(2\mu) \sum_{\alpha} \xi_{\alpha} + V\Psi = i\hbar u \quad (362)$$

$$\theta_{\xi\alpha} + \pi = \theta_V + \theta_{\Psi} = \theta_u + \pi/2 \quad (363)$$

or equivalently, Schrödinger's equation can be written approximately as

$$\begin{aligned} \hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{\text{vav}\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\sec \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial\Psi/\partial\theta_{\alpha}) \\ + V\Psi = i\hbar \sec \beta_{\Psi\Psi} \sin \beta_{tt}^{\Psi} t^{-1} \partial\Psi/\partial\theta_t^{\Psi} \end{aligned} \quad (364)$$

$$\beta_{\Psi\Psi} + \beta_{\text{vav}\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = \theta_V = \beta_{\Psi\Psi} - \theta_t^{\Psi} - \beta_{tt}^{\Psi} + \pi/2 \quad (365)$$

where  $\theta_{\Psi}$  has been subtracted from both sides of equation (365). The right hand sides of equations (362) and (364) must be real numbers. The  $\pi$  and  $\pi/2$  terms in equation (365) are necessary to pass to the limiting case of a slow quantum process in coherent space and coherent time.

The limiting form of equation (361) for the case of a slow quantum process in coherent space and coherent time is the following exact complex number Schrödinger equation

$$\hbar^2/(2\mu) \sum_{\alpha} 1/\bar{\alpha}^2 (F_{\Psi\alpha}^{\text{sc}} + jE_{\Psi\alpha}^{\text{sc}}) + \bar{V}\Psi = i\hbar/\bar{E} \partial\Psi/\partial\theta_t^{\Psi} \quad (366)$$

A combination of equations (360), (364) and (365) for the special case of a slow quantum process in coherent space and coherent time gives the approximate Schrödinger equation as

$$\hbar^2/(2\mu) \sum_{\alpha} \sec \beta_{\text{vav}\alpha}^{\text{sc}} \alpha^{-2} \partial^2\Psi/\partial\theta_{\alpha}^2 + V\Psi = i\hbar t^{-1} \partial\Psi/\partial\theta_t^{\Psi} \quad (367)$$

$$\beta_{\text{vav}\alpha}^{\text{sc}} - 2\theta_{\alpha} = \theta_V = -\theta_t^{\Psi} = \theta_E \quad (368)$$

Combining equations (92) through (94) with equation (367) gives the following approximate Schrödinger equation

$$\hbar^2/(2\mu) \sum_{\alpha} \alpha^{-2} [(E_{\Psi\alpha}^{\text{sc}})^2 + (F_{\Psi\alpha}^{\text{sc}})^2]^{1/2} + V\Psi = i\hbar t^{-1} \partial\Psi/\partial\theta_t^{\Psi} \quad (369)$$

Equation (369) also follows directly from equation (366). For the choice of the wave function for coherent time of the form

$$\begin{aligned} \Psi &= U \exp(-i\bar{E}t\theta_t^{\Psi}/\hbar) \\ &= U \exp(-iEt\theta_t^{\Psi}/\hbar) \end{aligned} \quad (370)$$

equation (366) becomes the stationary state time independent Schrödinger equation given in equation (203), while the choice of the wave function in equation (370) brings equation (369) into the form of the time independent Schrödinger equation (209).

**Case e. Ultrafast Quantum Process, Incoherent Space and Incoherent Time.**

This section considers the general form of the time dependent Schrödinger equation that is suitable to pass to the limiting case of an ultrafast quantum process in incoherent space and incoherent time which is described by

$$\beta_{\psi\psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \theta_{\tau}^{\psi} = 0 \quad \beta_{\tau\tau}^{\psi} = 0 \quad (371)$$

Equations (23), (59) and (325) give

$$\begin{aligned} & -\hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{\nu\alpha\nu\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi \partial\theta_{\psi}/\partial\alpha \partial\theta_{\nu\alpha}/\partial\alpha \exp(j\theta_{\xi\alpha}) \\ & + \bar{V}\bar{\psi} = i\hbar \csc \beta_{\psi\psi} \cos \beta_{\tau\tau}^{\psi} \psi \partial\theta_{\psi}/\partial\tau \exp(j\theta_u) \end{aligned} \quad (372)$$

Schrödinger's equation (372) can be represented approximately by the following two equations

$$\hbar^2/(2\mu) \sum_{\alpha} \xi_{\alpha} + V\psi = i\hbar u \quad (373)$$

$$\theta_{\xi\alpha} - \pi = \theta_V + \theta_{\psi} = \theta_u - \pi/2 \quad (374)$$

where the term  $\exp(j\theta_{\psi})$  has been factored out of equation (373). Equations (373) and (374) can be rewritten as the following approximate equations

$$\begin{aligned} & \hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{\nu\alpha\nu\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \partial\theta_{\psi}/\partial\alpha \partial\theta_{\nu\alpha}/\partial\alpha + V \\ & = i\hbar \csc \beta_{\psi\psi} \cos \beta_{\tau\tau}^{\psi} \partial\theta_{\psi}/\partial\tau \end{aligned} \quad (375)$$

$$\beta_{\psi\psi} + \beta_{\nu\alpha\nu\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) - \pi = \theta_V = \beta_{\psi\psi} - \theta_{\tau}^{\psi} - \beta_{\tau\tau}^{\psi} - \pi/2 \quad (376)$$

The  $-\pi$  and  $-\pi/2$  terms that appear in equations (374) and (376) are required to pass to the limiting case of an ultrafast quantum process in incoherent space and incoherent time.

In the special case of an ultrafast process in incoherent space and incoherent time equations (81) and (372) give the following exact complex number form of Schrödinger's equation

$$\hbar^2/(2\mu) \sum_{\alpha} (E_{\psi\alpha}^{ui} - jF_{\psi\alpha}^{ui}) + \bar{V} = ij\hbar \partial\theta_{\psi}/\partial\tau \quad (377)$$

where  $\bar{\psi}$  has been divided out of equation (372) to obtain equation (377). For

this special case equations (371), (375) and (376) give an approximate form of Schrödinger's equation as

$$\hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{v\alpha}^{ui} (\partial \theta_{\psi}/\partial \alpha)^2 + V = i\hbar \partial \theta_{\psi}/\partial t \quad (378)$$

$$\beta_{v\alpha}^{ui} - \pi/2 = \kappa_{\psi\alpha} = \theta_V = \theta_E = 0 \quad (379)$$

Equations (74) and (378) give the following approximate Schrödinger equation

$$\hbar^2/(2\mu) \sum_{\alpha} [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} + V = i\hbar \partial \theta_{\psi}/\partial t \quad (380)$$

Equation (380) also follows directly from equation (377). Equations (375), (377), (378) and (380) require that the stationary state for incoherent time is given by

$$\theta_{\psi} = \theta_U + \theta_{\psi t}^i \quad \theta_{\psi t}^i = -iEt/\hbar \quad (381)$$

which makes the right hand sides of equations (378) and (380) real numbers, and is the stationary state solution which when placed in equations (377) and (380) gives the time independent equations (243) and (246) respectively.

Case f. Ultrafast Quantum Process, Coherent Space and Incoherent Time.

This section examines the form of the time dependent Schrödinger equation that can be used to pass to the limit of an ultrafast quantum process in coherent space and incoherent time whose characteristics are

$$\beta_{\psi\psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad \theta_t^{\psi} = 0 \quad \beta_{tt}^{\psi} = 0 \quad (382)$$

Equations (23), (101) and (325) give the relevant time dependent Schrödinger equation as

$$\begin{aligned} & -\hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{v\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi/\alpha^2 \partial \theta_{\psi}/\partial \theta_{\alpha} \partial \theta_{v\alpha}/\partial \theta_{\alpha} \exp(j\theta_{\xi\alpha}) \\ & + \bar{V}\psi = i\hbar \csc \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \psi \partial \theta_{\psi}/\partial t \exp(j\theta_u) \end{aligned} \quad (383)$$

Equation (383) can be separated into the following two approximate Schrödinger equations by using equations (41), (115) and (116)

$$\hbar^2/(2\mu) \sum_{\alpha} \xi_{\alpha} + V\psi = i\hbar u \quad (384)$$

$$\theta_{\xi\alpha} + \pi = \theta_V + \theta_{\psi} = \theta_u - \pi/2 \quad (385)$$

which are equivalent to the following approximate Schrödinger equations

$$\begin{aligned} & \hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{v\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \alpha^{-2} \partial \theta_{\psi}/\partial \theta_{\alpha} \partial \theta_{v\alpha}/\partial \theta_{\alpha} + V \\ & = i\hbar \csc \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \partial \theta_{\psi}/\partial t \end{aligned} \quad (386)$$

$$\beta_{\psi\psi} + \beta_{\nu\alpha\nu\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = \theta_V = \beta_{\psi\psi} - \theta_t^{\psi} - \beta_{tt}^{\psi} - \pi/2 \quad (387)$$

where a factor  $\bar{\psi}$  has been divided out of equation (386). The right hand sides of equations (384) and (386) must be real numbers in internal space.

An exact complex number Schrödinger equation that is deduced from equation (383) in the limiting case of an ultrafast quantum process in coherent space and incoherent time is written as

$$\hbar^2/(2\mu) \sum_{\alpha} 1/\bar{\alpha}^2 (-E_{\psi\alpha}^{uc} + jF_{\psi\alpha}^{uc}) + \bar{V} = ij\hbar\partial\theta_{\psi}/\partial t \quad (388)$$

where a factor  $\bar{\psi}$  has been factored out to obtain equation (388). For this limiting case equations (382), (386) and (387) give the following approximate Schrödinger equations

$$\hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{\nu\alpha\nu\alpha}^{uc} \alpha^{-2} \partial\theta_{\psi}/\partial\theta_{\alpha} (\partial\theta_{\psi}/\partial\theta_{\alpha} - 1) + V = i\hbar\partial\theta_{\psi}/\partial t \quad (389)$$

$$\beta_{\nu\alpha\nu\alpha}^{uc} - 2\theta_{\alpha} + \pi/2 = \theta_V = \theta_E = 0 \quad (390)$$

Equations (389) and (390) can be rewritten using equations (126) and (128) resulting in the approximate Schrödinger equations given by

$$\hbar^2/(2\mu) \sum_{\alpha} \alpha^{-2} [(E_{\psi\alpha}^{uc})^2 + (F_{\psi\alpha}^{uc})^2]^{1/2} + V = i\hbar\partial\theta_{\psi}/\partial t \quad (391)$$

$$\delta_{\psi\alpha} - 2\theta_{\alpha} = \theta_V = \theta_E = 0 \quad (392)$$

where  $E_{\psi\alpha}^{uc}$  and  $F_{\psi\alpha}^{uc}$  are given by equations (121) and (122), and where the right hand sides of equations (389) and (391) must be real numbers in internal space. The approximate Schrödinger equations (391) and (392) can also be obtained directly from equation (388). Schrödinger's equation (391) has a stationary state solution for incoherent time which is of the form

$$\theta_{\psi} = \theta_U + \theta_{\psi t}^i \quad \theta_{\psi t}^i = -iEt/\hbar \quad (393)$$

and equations (388) and (391) become the time independent Schrödinger equations (282) and (288) respectively.

Case g. Ultrafast Quantum Process, Incoherent Space and Coherent Time.

This section considers the time dependent Schrödinger equation that can be used to obtain the limiting form of the equation for the case of an ultrafast quantum process in incoherent space and coherent time which is described by

$$\beta_{\psi\psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \beta_{tt}^{\psi} = \pi/2 \quad (394)$$

Equations (25), (59) and (325) give Schrödinger's equation as

$$\begin{aligned}
& - \hbar^2 / (2\mu) \sum_{\alpha} \csc \beta_{\text{vava}} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\Psi\Psi} \Psi \partial \theta_{\Psi} / \partial \alpha \partial \theta_{\text{va}} / \partial \alpha \exp(j\theta_{\xi\alpha}) \quad (395) \\
& + \bar{V}\bar{\Psi} = i\hbar \csc \beta_{\Psi\Psi} \sin \beta_{\text{tt}}^{\Psi} \Psi / t \partial \theta_{\Psi} / \partial \theta_{\text{t}}^{\Psi} \exp(j\theta_{\text{u}})
\end{aligned}$$

Equation (395) is represented by the following approximate scalar Schrödinger equations by using equations (16), (64) and (65)

$$\hbar^2 / (2\mu) \sum_{\alpha} \xi_{\alpha} + V\Psi = i\hbar u \quad (396)$$

$$\theta_{\xi\alpha} - \pi = \theta_V + \theta_{\Psi} = \theta_u \quad (397)$$

which are written equivalently as the following approximate Schrödinger equations by using equations (25), (26), (59) and (68)

$$\hbar^2 / (2\mu) \sum_{\alpha} \csc \beta_{\text{vava}} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\Psi\Psi} \partial \theta_{\Psi} / \partial \alpha \partial \theta_{\text{va}} / \partial \alpha + V \quad (398)$$

$$= i\hbar \csc \beta_{\Psi\Psi} \sin \beta_{\text{tt}}^{\Psi} t^{-1} \partial \theta_{\Psi} / \partial \theta_{\text{t}}^{\Psi}$$

$$\beta_{\Psi\Psi} + \beta_{\text{vava}} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) - \pi = \theta_V = \beta_{\Psi\Psi} - \theta_{\text{t}}^{\Psi} - \beta_{\text{tt}}^{\Psi} \quad (399)$$

where a factor  $\Psi$  has been divided out of equation (398) and the angle  $\theta_{\Psi}$  has been subtracted from equation (397) to obtain equation (399).

For the case of an ultrafast quantum process in incoherent space and coherent time the limiting form of equation (395) can be obtained from equations (48), (81) and (324) to be the following exact Schrödinger equation

$$\hbar^2 / (2\mu) \sum_{\alpha} (E_{\Psi\alpha}^{\text{ui}} - jF_{\Psi\alpha}^{\text{ui}}) + \bar{V} = i\hbar/t \partial \theta_{\Psi} / \partial \theta_{\text{t}}^{\Psi} \quad (400)$$

where, a factor  $\bar{\Psi}$  has been divided out of equation (400). Combining equations (394), (398) and (399) gives the following approximate Schrödinger equations for the special case of an ultrafast quantum process that occurs in incoherent space and coherent time

$$\hbar^2 / (2\mu) \sum_{\alpha} \csc \beta_{\text{vava}}^{\text{ui}} (\partial \theta_{\Psi} / \partial \alpha)^2 + V = i\hbar t^{-1} \partial \theta_{\Psi} / \partial \theta_{\text{t}}^{\Psi} \quad (401)$$

$$\beta_{\text{vava}}^{\text{ui}} - \pi/2 = \theta_V = \theta_E = -\theta_{\text{t}}^{\Psi} \quad (402)$$

Using equations (74) and (76) allows equations (401) and (402) to be written as the following approximate Schrödinger equations

$$\hbar^2 / (2\mu) \sum_{\alpha} [(E_{\Psi\alpha}^{\text{ui}})^2 + (F_{\Psi\alpha}^{\text{ui}})^2]^{1/2} + V = i\hbar t^{-1} \partial \theta_{\Psi} / \partial \theta_{\text{t}}^{\Psi} \quad (403)$$

$$\kappa_{\Psi\alpha} = \theta_V = \theta_E = -\theta_{\text{t}}^{\Psi} \quad (404)$$

The approximate Schrödinger equations in equations (403) and (404) can be obtained directly from the exact equation (400). Schrödinger's time dependent equations (400), (401) and (403) for coherent time have a stationary state solution of the form

$$\theta_{\Psi} = \theta_U + \theta_{\Psi t}^c \quad (405)$$

$$\theta_{\Psi t}^c = -i\bar{E}t\theta_t^{\Psi}/\hbar = -iEt\theta_t^{\Psi}/\hbar \quad (406)$$

where the following reality condition has been used

$$\bar{E}t = Et \quad (407)$$

which is valid for wave propagation with constant  $E$ .<sup>19</sup> It should be added that for wave propagation with constant wave number  $k_{\alpha}$  and frequency  $\omega$  the following conditions are also valid<sup>19</sup>

$$\bar{k}_{\alpha} \bar{\alpha} = k_{\alpha} \alpha \quad \bar{\omega}t = \omega t \quad (408)$$

where  $\alpha = x, y, z$ . Substituting equations (405) and (406) into (400) and (403) yields the time independent stationary state equations (243) and (246) respectively.

Case h. Ultrafast Quantum Process, Coherent Space and Coherent Time.

Finally, this section develops the form of the time dependent Schrödinger equation that can be used to pass to the limit of an ultrafast quantum process in coherent space and coherent time which is characterized by

$$\beta_{\Psi\Psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad \beta_{tt}^{\Psi} = \pi/2 \quad (409)$$

Equations (35), (101) and (325) give the proper form of Schrödinger's equation as

$$\begin{aligned} & -\hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{\alpha\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{\Psi\Psi} \Psi/\alpha^2 \partial\theta_{\Psi}/\partial\theta_{\alpha} \partial\theta_{\alpha}/\partial\theta_{\alpha} \exp(j\theta_{\alpha}) \\ & + \bar{V}\bar{\Psi} = i\hbar \csc \beta_{\Psi\Psi} \sin \beta_{tt}^{\Psi} \Psi/t \partial\theta_{\Psi}/\partial\theta_t^{\Psi} \exp(j\theta_u) \end{aligned} \quad (410)$$

As before, equation (410) can be separated into two approximate scalar Schrödinger equations

$$\hbar^2/(2\mu) \sum_{\alpha} \xi_{\alpha} + V\Psi = i\hbar u \quad (411)$$

$$\theta_{\xi\alpha} + \pi = \theta_V + \theta_{\Psi} = \theta_u \quad (412)$$

which can be written equivalently as the following approximate Schrödinger equations



$$\begin{aligned} & \hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{\alpha\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{\Psi\Psi} \alpha^{-2} \partial \theta_{\Psi}/\partial \theta_{\alpha} \partial \theta_{\Psi\alpha}/\partial \theta_{\alpha} + V \\ & = i\hbar \csc \beta_{\Psi\Psi} \sin \beta_{\Psi\Psi} t^{-1} \partial \theta_{\Psi}/\partial \theta_t^{\Psi} \end{aligned} \quad (413)$$

$$\beta_{\Psi\Psi} + \beta_{\alpha\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = \theta_V = \beta_{\Psi\Psi} - \theta_t^{\Psi} - \beta_{\Psi\Psi} \quad (414)$$

where  $\theta_{\alpha}$  and  $\theta_t^{\Psi}$  are variables.

The limiting form of equation (410) for the special case of an ultrafast quantum process in coherent space and coherent time can be obtained using equations (48), (135) and (324) and yields an exact Schrödinger equation for this case

$$\hbar^2/(2\mu) \sum_{\alpha} 1/\alpha^2 (-E_{\Psi\alpha}^{uc} + jF_{\Psi\alpha}^{uc}) + \bar{V} = i\hbar/t \partial \theta_{\Psi}/\partial \theta_t^{\Psi} \quad (415)$$

where  $\bar{\Psi}$  has been divided out of equation (410) to obtain equation (415). In the special case of an ultrafast quantum process in coherent space and coherent time, equations (409), (413) and (414) give the approximate Schrödinger equations as

$$\hbar^2/(2\mu) \sum_{\alpha} \csc \beta_{\alpha\alpha}^{uc} \alpha^{-2} \partial \theta_{\Psi}/\partial \theta_{\alpha} (\partial \theta_{\Psi}/\partial \theta_{\alpha} - 1) + V = i\hbar t^{-1} \partial \theta_{\Psi}/\partial \theta_t^{\Psi} \quad (416)$$

$$\beta_{\alpha\alpha}^{uc} - 2\theta_{\alpha} + \pi/2 = \theta_V = \theta_E = -\theta_t^{\Psi} \quad (417)$$

These two equations can be rewritten using equations (126) and (128) as the following approximate Schrödinger equations

$$\hbar^2/(2\mu) \sum_{\alpha} \alpha^{-2} [(E_{\Psi\alpha}^{uc})^2 + (F_{\Psi\alpha}^{uc})^2]^{1/2} + V = i\hbar t^{-1} \partial \theta_{\Psi}/\partial \theta_t^{\Psi} \quad (418)$$

$$\delta_{\Psi\alpha} - 2\theta_{\alpha} = \theta_V = \theta_E = -\theta_t^{\Psi} \quad (419)$$

The total internal phase angle of the wave function is written as in equations (405) and (406). The approximate equations (418) and (419) can also be obtained directly from the exact Schrödinger equation (415). Equations (415) and (418) have stationary state solutions of the form given by equations (405) and (406) which reduce equations (415) and (418) to their time independent forms given by equations (282) and (288) respectively.

**4. DIRAC EQUATION IN BROKEN SYMMETRY SPACETIME.** This section investigates slow and ultrafast relativistic quantum processes in coherent and incoherent spacetime. The broken symmetry form of the Dirac equation is written as the following complex number generalization of the standard Dirac equation<sup>28,29</sup>

$$-i\hbar \gamma_{\mu} \partial \bar{\Psi}/\partial \bar{x}_{\mu} + m\bar{\Psi} = 0 \quad (420)$$

where  $\gamma_{\mu}$  = Dirac matrices,  $\bar{\Psi}$  = complex number four component Dirac spinor which

is now a set of four complex numbers in an internal space, and where the complex number space and time coordinates are designated by

$$\bar{x}_\mu = x_\mu \exp(j\theta_{x\mu}) \quad (421)$$

where  $\mu = 0, 1, 2, 3$  and  $x_0 = ct$ ,  $x_1 = x$ ,  $x_2 = y$  and  $x_3 = z$ . The Dirac equation (420) can be rewritten as follows

$$-i\hbar\gamma_\mu \bar{v}^\mu + m\bar{\Psi} = 0 \quad (422)$$

where

$$\bar{v}^\mu = v^\mu \exp(j\theta_{v\mu}) = \partial\bar{\Psi}/\partial\bar{x}_\mu \quad (423)$$

where

$$v^\mu = \sec \beta_{\Psi\Psi} \cos \beta_{x\mu x\mu} \partial\Psi/\partial x_\mu \quad (424)$$

$$= \csc \beta_{\Psi\Psi} \cos \beta_{x\mu x\mu} \Psi \partial\theta_\Psi/\partial x_\mu \quad (425)$$

$$= \sec \beta_{\Psi\Psi} \sin \beta_{x\mu x\mu} x_\mu^{-1} \partial\Psi/\partial\theta_{x\mu} \quad (426)$$

$$= \csc \beta_{\Psi\Psi} \sin \beta_{x\mu x\mu} \Psi/x_\mu \partial\theta_\Psi/\partial\theta_{x\mu} \quad (427)$$

where  $\beta_{\Psi\Psi}$  is given by equation (14) and where

$$\theta_{v\mu} = \theta_\Psi + \beta_{\Psi\Psi} - \theta_{x\mu} - \beta_{x\mu x\mu} \quad (428)$$

$$\tan \beta_{x\mu x\mu} = x_\mu \partial\theta_{x\mu}/\partial x_\mu$$

The component parts of the Dirac equation (422) can be written as

$$-i\hbar\gamma_\mu v^\mu \cos \theta_{v\mu} + m\Psi \cos \theta_\Psi = 0 \quad (430)$$

$$-i\hbar\gamma_\mu v^\mu \sin \theta_{v\mu} + m\Psi \sin \theta_\Psi = 0 \quad (431)$$

where the mass  $m$  is taken to be a scalar. An approximate representation of equation (422) follows from the assumption that the phase angles of each term in equation (422) are equal, with the result that the Dirac equation is written as

$$-i\hbar\gamma_\mu v^\mu + m\Psi = 0 \quad (432)$$

$$\theta_{v\mu} = \theta_\Psi \quad (433)$$

Equations (428) and (433) give for this approximation

$$\beta_{\Psi\Psi} - \theta_{x\mu} - \beta_{x\mu x\mu} = 0 \quad (434)$$

Equation (434) is a differential equation relating  $\theta_{\Psi}$  and  $\theta_{x\mu}$  as can be seen from equations (14) and (429). For the case of a slow quantum process (incoherent wave function) in incoherent spacetime, all internal phase angles are set to zero and equations (420), (424) and (432) reduce to the standard Dirac equation

$$-i\hbar\gamma_{\mu}\partial\Psi/\partial x_{\mu} + m\Psi = 0 \quad (435)$$

where  $\Psi$  is the standard Dirac four component spinor.

For coherent spacetime coordinate variation and an arbitrary variation of the wave function, equation (420) becomes the Dirac equation

$$ij\hbar/\bar{x}_{\mu}\gamma_{\mu}\partial\bar{\Psi}/\partial\theta_{x\mu} + m\bar{\Psi} = 0 \quad (436)$$

For the special case of an ultrafast quantum process in coherent spacetime equations (420) or (436) give the following Dirac equation

$$(-i\hbar/\bar{x}_{\mu}\gamma_{\mu}\partial\theta_{\Psi}/\partial\theta_{x\mu} + m)\bar{\Psi} = 0 \quad (437)$$

Equation (437) is a matrix equation where

$$\bar{\Psi} = \begin{pmatrix} \bar{\Psi}_1 \\ \bar{\Psi}_2 \\ \bar{\Psi}_3 \\ \bar{\Psi}_4 \end{pmatrix} \quad \Psi = \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \Psi_3 \\ \Psi_4 \end{pmatrix} \quad (438)$$

and where

$$\theta_{\Psi} = \begin{pmatrix} \theta_{\Psi 1} & 0 & 0 & 0 \\ 0 & \theta_{\Psi 2} & 0 & 0 \\ 0 & 0 & \theta_{\Psi 3} & 0 \\ 0 & 0 & 0 & \theta_{\Psi 4} \end{pmatrix} \quad (439)$$

$$\partial\theta_{\Psi}/\partial\theta_{x\mu} = \begin{pmatrix} \partial\theta_{\Psi 1}/\partial\theta_{x\mu} & 0 & 0 & 0 \\ 0 & \partial\theta_{\Psi 2}/\partial\theta_{x\mu} & 0 & 0 \\ 0 & 0 & \partial\theta_{\Psi 3}/\partial\theta_{x\mu} & 0 \\ 0 & 0 & 0 & \partial\theta_{\Psi 4}/\partial\theta_{x\mu} \end{pmatrix} \quad (440)$$

where  $\mu = 0,1,2,3$  . The Dirac equation (437) for an ultrafast quantum process in coherent spacetime can also be deduced from equations (422), (427) and (428) by setting  $\beta_{\Psi\Psi} = \pi/2$  and  $\beta_{x_{\mu}x_{\mu}} = \pi/2$  which gives

$$\bar{v}^{\mu} = \bar{\Psi}/x_{\mu} \partial\theta_{\Psi}/\partial\theta_{x_{\mu}} \quad (441)$$

$$v^{\mu} = \Psi/x_{\mu} \partial\theta_{\Psi}/\partial\theta_{x_{\mu}} \quad (442)$$

$$\theta_{v\mu} = \theta_{\Psi} - \theta_{x_{\mu}} \quad (443)$$

The real and imaginary parts of equation (437) can be written as the following matrix equations

$$[-i\hbar/x_{\mu} \gamma_{\mu} \partial\theta_{\Psi}/\partial\theta_{x_{\mu}} \cos(\theta_{\Psi} - \theta_{x_{\mu}}) + m \cos \theta_{\Psi}] \Psi = 0 \quad (444)$$

$$[-i\hbar/x_{\mu} \gamma_{\mu} \partial\theta_{\Psi}/\partial\theta_{x_{\mu}} \sin(\theta_{\Psi} - \theta_{x_{\mu}}) + m \sin \theta_{\Psi}] \Psi = 0 \quad (445)$$

where  $\cos(\theta_{\Psi} - \theta_{x_{\mu}})$  ,  $\sin(\theta_{\Psi} - \theta_{x_{\mu}})$  ,  $\cos \theta_{\Psi}$  and  $\sin \theta_{\Psi}$  are 4x4 diagonal matrices.

Approximate matrix equations that result from the assumption that  $\theta_{x_{\mu}} \sim 0$  in equations (444) and (445) are the following

$$(-i\hbar/x_{\mu} \gamma_{\mu} \partial\theta_{\Psi}/\partial\theta_{x_{\mu}} + m) \Psi_R \sim 0 \quad (446)$$

$$(-i\hbar/x_{\mu} \gamma_{\mu} \partial\theta_{\Psi}/\partial\theta_{x_{\mu}} + m) \Psi_I \sim 0 \quad (447)$$

where the column matrices  $\Psi_R$  and  $\Psi_I$  are given by the following matrix equations

$$\Psi_R = \cos \theta_{\Psi} \Psi = \begin{pmatrix} \Psi_1 \cos \theta_{\Psi 1} \\ \Psi_2 \cos \theta_{\Psi 2} \\ \Psi_3 \cos \theta_{\Psi 3} \\ \Psi_4 \cos \theta_{\Psi 4} \end{pmatrix} \quad (448)$$

$$\Psi_I = \sin \theta_{\Psi} \Psi = \begin{pmatrix} \Psi_1 \sin \theta_{\Psi 1} \\ \Psi_2 \sin \theta_{\Psi 2} \\ \Psi_3 \sin \theta_{\Psi 3} \\ \Psi_4 \sin \theta_{\Psi 4} \end{pmatrix} \quad (449)$$

Choose the following linear solution for the diagonal matrix  $\theta_{\Psi}$

$$\theta_{\Psi} = \sum_{\mu} b_{\mu}^{\Psi} \theta_{x_{\mu}} \quad (450)$$

where  $b_{\mu}^{\Psi}$  are four constant diagonal matrices given by

$$b_{\mu}^{\Psi} = \begin{pmatrix} b_{\mu}^{\Psi 1} & 0 & 0 & 0 \\ 0 & b_{\mu}^{\Psi 2} & 0 & 0 \\ 0 & 0 & b_{\mu}^{\Psi 3} & 0 \\ 0 & 0 & 0 & b_{\mu}^{\Psi 4} \end{pmatrix} \quad (451)$$

where  $\mu = 0, 1, 2, 3$  ; so that  $b_{\mu}^{\Psi}$  represents sixteen quantities. The component representation of the matrix equation (450) is written as

$$\theta_{\Psi \nu} = \sum_{\mu} b_{\mu}^{\Psi \nu} \theta_{x \mu} \quad (452)$$

where

$$b_{\mu}^{\Psi \nu} = \partial \theta_{\Psi \nu} / \partial \theta_{x \mu} \quad (453)$$

for  $\mu = 0, 1, 2, 3$  and  $\nu = 0, 1, 2, 3$  . Then Dirac's equations (437), (446) and (447) can be written as

$$(-i\hbar/\bar{x}_{\mu} \gamma_{\mu} b_{\mu}^{\Psi} + m)\bar{\Psi} = 0 \quad (454)$$

$$(-i\hbar/x_{\mu} \gamma_{\mu} b_{\mu}^{\Psi} + m)\Psi_R \sim 0 \quad (455)$$

$$(-i\hbar/x_{\mu} \gamma_{\mu} b_{\mu}^{\Psi} + m)\Psi_I \sim 0 \quad (456)$$

which are valid for ultrafast relativistic quantum processes in coherent spacetime.

Finally it should be pointed out that from equations (422) and (423) it follows that the Dirac equation for an ultrafast quantum process in incoherent spacetime is given by

$$(-ij\hbar\gamma_{\mu} \partial \theta_{\Psi} / \partial x_{\mu} + m)\bar{\Psi} = 0 \quad (457)$$

where  $\partial \theta_{\Psi} / \partial x_{\mu}$  are four diagonal matrices.

$$\partial \theta_{\Psi} / \partial x_{\mu} = \begin{pmatrix} \partial \theta_{\Psi 1} / \partial x_{\mu} & 0 & 0 & 0 \\ 0 & \partial \theta_{\Psi 2} / \partial x_{\mu} & 0 & 0 \\ 0 & 0 & \partial \theta_{\Psi 3} / \partial x_{\mu} & 0 \\ 0 & 0 & 0 & \partial \theta_{\Psi 4} / \partial x_{\mu} \end{pmatrix} \quad (458)$$

for  $\mu = 0, 1, 2, 3$ . The real and imaginary parts of equation (457) are given by

$$(i\hbar\gamma_{\mu} \partial\theta_{\psi}/\partial x_{\mu} \sin \theta_{\psi} + m \cos \theta_{\psi})\psi = 0 \quad (459)$$

$$(-i\hbar\gamma_{\mu} \partial\theta_{\psi}/\partial x_{\mu} \cos \theta_{\psi} + m \sin \theta_{\psi})\psi = 0 \quad (460)$$

where  $\cos \theta_{\psi}$  and  $\sin \theta_{\psi}$  are diagonal matrices in analogy to equation (439).

**5. CONCLUSION.** Quantum mechanics can be formulated for the case where the wave function and the space and time coordinates have broken internal symmetries which require that these quantities be written as complex numbers in an internal space. This allows the possibility of slow and ultrafast quantum processes occurring in spacetime with coherent and incoherent variation of coordinates. For an ultrafast quantum process the complex number wave function rotates in an internal space, while for a slow quantum process the wave function changes in magnitude. Spacetime coordinates can likewise change coherently by a rotation in an internal space or incoherently by a change in the magnitudes of the space and time coordinates. The Dirac equation and the Schrödinger time dependent and time independent equations are formulated for slow and ultrafast quantum processes that can occur in coherent or incoherent space and time. Ultrafast processes are important to science and engineering and may possibly be the basis of developing energy sources such as the nuclear rocket engine.

#### ACKNOWLEDGEMENT

Elizabeth K. Klein typed and edited this paper, for which I am very grateful.

#### REFERENCES

1. Shapiro, S. L., editor, Ultrashort Light Pulses, Springer-Verlag, New York, 1984.
2. Auston, D. H., "Probing Semiconductors with Femtosecond Pulses," *Physics Today*, pg. 46, February 1990.
3. Zewail, A. H., "Laser Femtochemistry," *Science*, Vol. 242, pg. 1645, 23 December 1988.
4. Gruebele, M. and Zewail, A. H., "Ultrafast Reaction Dynamics," *Physics Today*, pg. 24, May 1990.
5. Stoutland, P. O., Dyer, R. B. and Woodruff, W. H., "Ultrafast Infrared Spectroscopy," *Science*, Vol. 257, pg. 1913, 25 September 1992.
6. Antonsen, T. M. and Mora, P., "Self-Focusing and Raman Scattering of Laser Pulses in Tenuous Plasmas," *Phys. Rev. Lett.*, Vol. 69, pg. 2204, 12 October 1992.
7. Baumert, R., Röttgermann, C., Rothenfusser, C., Thalweiser, R., Weiss, V., and Gerber, G., "Femtosecond Probing of Sodium Cluster Ion  $\text{Na}_n^+$  Fragmentation," *Physics. Rev. Lett.*, Vol. 69, pg. 1512, 7 September 1992.

8. Wright, O. B. and Kawashima, K., "Coherent Phonon Detection from Ultrafast Surface Vibrations," Phys. Rev. Lett., Vol. 69, pg. 1668, 14 September 1992.
9. Snoke, D. W., Rühle, W. W., Lu, Y. C. and Bauser, E., "Nonthermalized Distribution of Electrons on Picosecond Time Scale in GaAs," Phys. Rev. Lett., Vol. 68, pg. 990, 17 February 1992.
10. Noordam, L. D., Stapelfeldt, H., Duncan, D. I. and Gallagher, T. F., "Redistribution of Rydberg States by Intense Picosecond Pulses," Phys. Rev. Lett., Vol. 68, pg. 1496, 9 March 1992.
11. Judson, R. S. and Rabitz, H., "Teaching Lasers to Control Molecules," Phys. Rev. Lett., Vol. 68, pg. 1500, 9 March 1992.
12. Kmetec, J. D., Gordon III, C. L., Macklin, J. J., Lemoff, B. E., Brown, G. S. and Harris, S. E., "MeV X-Ray Generation with a Femtosecond Laser," Phys. Rev. Lett., Vol. 68, pg. 1527, 9 March 1992.
13. Nibbering, E. T. J., Wiersma, D. A. and Duppen, K., "Ultrafast Nonlinear Spectroscopy with Chirped Optical Pulses," Phys. Rev. Lett., Vol. 68, pg. 514, 24 January 1992.
14. Steinmüller-Nethl, D., Höpfel, R. A., Gornik, E., Leitner, A. and Aussenegg, F. R., "Femtosecond Relaxation of Localized Plasma Excitations in Ag Islands," Phys. Rev. Lett., Vol. 68, pg. 389, 20 January 1992.
15. Potter, E. D., Herek, J. L., Pedersen, S., Liu, Q. and Zewail, A. H., "Femtosecond Laser Control of a Chemical Reaction," Nature, Vol. 355, pg. 66, 2 January 1992.
16. Hu, B. B., Zhang, X. C. and Auston, D. H., "Terahertz Radiation Induced by a Subband-Gap Femtosecond Optical Excitation of GaAs," Phys. Rev. Lett., Vol. 67, pg. 2709, 4 November 1991.
17. Wood, W. M., Siders, C. W. and Downer, M. C., "Measurement of Femtosecond Ionization Dynamics of Atmospheric Density Gases by Spectral Blueshifting," Phys. Rev. Lett., Vol. 67, pg. 3523, 16 December 1991.
18. Weiss, R. A., Relativistic Thermodynamics, Vols. 1&2, K&W Publications, Vicksburg, MS, 1976.
19. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
20. Weiss, R. A., Clean Fission, K&W Publications, Vicksburg, MS, 1992.
21. Weiss, R. A., "Slow and Ultrafast Wave Propagation Processes," article in Clean Fission, K&W Publications, Vicksburg, 1992.
22. Atkins, P. W., Quanta, Oxford University Press, New York, 1991.

23. Merzbacher, E., Quantum Mechanics, John Wiley, New York, 1961.
24. Kursunoglu, B., Modern Quantum Theory, W. H. Freeman, San Francisco, 1962.
25. Pauling, L. and Wilson, E. B., Introduction to Quantum Mechanics, McGraw-Hill, New York, 1935.
26. Bethe, H. A. and Jackiw, R., Intermediate Quantum Mechanics, Benjamin, New York, 1968.
27. Weiss, R. A., "Quantum Mechanics and the Broken Symmetry of Space and Time," article in Clean Fission, K&W Publications, Vicksburg, 1992.
28. Greiner, W., Müller, B. and Rafelski, J., Quantum Electrodynamics of Strong Fields, Springer-Verlag, New York, 1985.
29. Itzykson, C. and Zuber, J., Quantum Field Theory, McGraw-Hill, New York, 1980.



## QUATERNARY FISSION OF $\gamma$ RAY COOLED ACTINIDES

Richard A. Weiss

U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** The basic concepts needed for the design of clean fission nuclear reactors that use actinide elements such as  $^{235}\text{U}$  and  $^{239}\text{Pu}$  as fuels are developed in this paper. In the presence of an electromagnetic field all nuclei exhibit a broken internal symmetry that can be described by requiring the atomic number and atomic mass number to be complex numbers whose magnitudes are functions of the internal phase angles of the complex numbers. The form of these functions depends on the value of the fissility parameter of a nucleus. For subactinide nuclei with fissility parameters less than unity the measured atomic number and measured atomic mass number are not exactly integers, while for actinide nuclei whose fissility parameters are greater than unity the measured atomic number and atomic mass number must be integers. This basic difference in the countability of nucleons in subactinide nuclei and actinide nuclei yields two different representations of the complex number form of the liquid drop type of nuclear mass formula. For actinide nuclei in an external electromagnetic field the complex number Bohr-Wheeler fission condition predicts that above a critical  $\gamma$  ray intensity, that is determined by the value of the fissility parameter, the thermal neutron induced binary fission mode is suppressed. The alternate mode of quaternary fission due to the thermal neutron induced binary fission of the two component subactinide lobes of an actinide nucleus into four fission product nuclei is catalyzed by the presence of a  $\gamma$  ray field. A mathematical expression is derived in terms of the fissility parameter that gives the value of the critical electromagnetic field strength of the external  $\gamma$  ray field required to cool and suppress thermal neutron induced binary fission of the actinides. The nuclear fission waste products of the  $\gamma$  ray catalyzed thermal neutron induced quaternary fission of the actinide elements will be low level radionuclides, and represents a way of generating clean fission nuclear power. Clean fission nuclear reactor designs are considered that may be used for submarines and nuclear rocket engines.

**1. INTRODUCTION.** Nuclear fission reactors may yet be the power source that will drive the industrial nations in the future. At present, however, nuclear reactors are a potentially dangerous way of generating power. This is due first to the possibility of nuclear accidents which can spread radioactive fission products such as  $^{90}\text{Sr}$  into the environment, and secondly to the growing problem of the storage of radioactive wastes.<sup>1-7</sup> Power generated from nuclear fission was immediately recognized as an alternative to the energy generated by burning fossil fuels.<sup>8,9</sup> The use of fossil fuels causes air pollution which results in acid rain, various human diseases, and may even contribute to global warming.<sup>10-14</sup> Nuclear fission power is used in countries like Japan and France to generate a significant portion of their energy needs because these countries have no natural fossil fuel resources and are therefore forced by economic necessity to utilize nuclear fission reactors for generating electricity. No major nuclear accident has occurred in these countries, but serious nuclear acci-

dents have occurred in the United States and Ukraine. Therefore scientists and engineers are now looking for alternative energy sources which are safe and non-polluting.

But what are the alternatives to fossil fuels and conventional nuclear fission power? A glance at history shows that people have always used their minds to bend nature to their advantage.<sup>15-18</sup> And today the search for new energy sources continues in various directions such as solar power, wind power, geothermal energy, biomass energy and nuclear fusion power.<sup>19-25</sup> These new energy sources are nowhere near the point of scientific and engineering development where they can be used as alternative power sources to fossil fuels or nuclear fission power. Nuclear fusion, the power source of the stars, remains etherial after many years of research and may ultimately be a chimera. In view of this it may be wise to take a second look at nuclear fission power to see if new concepts can be developed which will allow nuclear reactors to be designed for safe and environmentally benign operation. A clean fission nuclear reactor core design concept has been proposed that uses  $\gamma$  ray catalyzed thermal neutron induced binary fission of subactinide elements.<sup>26</sup> In the present paper a  $\gamma$  ray catalyzed thermal neutron induced quaternary fission nuclear reactor is investigated that uses actinide elements as a source of clean fission power. In this case the  $\gamma$  rays are used to suppress binary fission and catalyze quaternary fission in the actinides whose reaction product nuclei will be relatively light weight nuclei having only low level beta decays.

The use of  $\gamma$  rays for the suppression of binary fission and the enhancement of quaternary fission in the actinides is possible because atomic nuclei are systems of matter that can exhibit broken internal symmetries which are ultimately related to the broken internal symmetries of spacetime.<sup>27</sup> The broken symmetries of spacetime are described by representing time and space coordinates as complex numbers in the following way<sup>27</sup>

$$\bar{t} = t \exp(j\theta_t) \quad (1)$$

and for cartesian coordinates

$$\bar{x} = x \exp(j\theta_x) \quad \bar{y} = y \exp(j\theta_y) \quad \bar{z} = z \exp(j\theta_z) \quad (2)$$

while for spherical polar coordinates

$$\bar{r} = r \exp(j\theta_r) \quad \bar{\phi} = \phi \exp(j\theta_\phi) \quad \bar{\psi} = \psi \exp(j\theta_\psi) \quad (3)$$

Corresponding to the complex number azimuthal angle is the following complex magnetic quantum number<sup>27</sup>

$$\bar{M} = M \exp(j\theta_M) = m \cos \theta_\phi \exp(-j\theta_\phi) \quad (4)$$

where  $m = 0, \pm 1, \pm 2, \pm 3, \dots$  is the standard magnetic quantum number. Therefore<sup>27</sup>

$$M = m \cos \theta_\phi \quad \theta_M = -\theta_\phi \quad (5)$$

It is often convenient to define the following positive magnetic quantum number<sup>27</sup>

$$\bar{M}' = M' \exp(j\theta_M') = |m| \cos \theta_\phi \exp(-j\theta_\phi) \quad (6)$$

$$M' = |m| \cos \theta_\phi \quad \theta_M' = -\theta_\phi \quad (7)$$

For  $\theta_\phi = 0$  these quantities reduce to the standard concepts.

It is assumed that the integer quantum numbers Z, N and A are analogous to the magnetic quantum number  $|m|$  and occur in a solution to an azimuthal portion of a wave equation in internal space.<sup>27</sup> Then by an argument similar to that for the complex number magnetic quantum number it follows that the atomic number, neutron number and atomic mass number are complex numbers in an internal space and can be written in a form similar to that of the complex magnetic quantum number  $\bar{M}'$  in equation (6) as follows for subactinide nuclei located in an external field<sup>26</sup>

$$\bar{z} = z \exp(j\theta_z) = Z \cos \theta_z \exp(j\theta_z) \quad (8)$$

$$\bar{n} = n \exp(j\theta_n) = N \cos \theta_n \exp(j\theta_n) \quad (9)$$

$$\bar{a} = a \exp(j\theta_a) = A \cos \theta_a \exp(j\theta_a) \quad (10)$$

The complex number azimuthal angles corresponding to the quantum numbers  $\bar{z}$ ,  $\bar{n}$  and  $\bar{a}$  are written, in analogy to the azimuthal angle of real space given in equation (3), as follows

$$\bar{\phi}_z = \phi_z \exp(j\theta_{\phi z}) \quad \bar{\phi}_n = \phi_n \exp(j\theta_{\phi n}) \quad \bar{\phi}_a = \phi_a \exp(j\theta_{\phi a}) \quad (11)$$

where  $\theta_{\phi z}$ ,  $\theta_{\phi n}$  and  $\theta_{\phi a}$  are the internal phase angles of the complex number azimuthal angles in the internal space of nucleons, so that finally for subactinide nuclei in an external field<sup>26</sup>

$$z = Z \cos \theta_z \quad n = N \cos \theta_n \quad a = A \cos \theta_a \quad (12)$$

$$\theta_z = -\theta_{\phi z} \quad \theta_n = -\theta_{\phi n} \quad \theta_a = -\theta_{\phi a} \quad (13)$$

The real and imaginary parts of equations (8) through (10) are given for subactinide nuclei in an external field by<sup>26</sup>

$$z_R = Z \cos^2 \theta_z = Z \cos^2 \theta_{\phi z} \quad (14)$$

$$n_R = N \cos^2 \theta_n = N \cos^2 \theta_{\phi n} \quad (15)$$

$$a_R = A \cos^2 \theta_a = A \cos^2 \theta_{\phi a} \quad (16)$$

$$z_I = Z \cos \theta_z \sin \theta_z = -Z \cos \theta_{\phi z} \sin \theta_{\phi z} \quad (17)$$

$$n_I = N \cos \theta_n \sin \theta_n = -N \cos \theta_{\phi n} \sin \theta_{\phi n} \quad (18)$$

$$a_I = A \cos \theta_a \sin \theta_a = -A \cos \theta_{\phi a} \sin \theta_{\phi a} \quad (19)$$

The measured atomic number, neutron number and atomic mass number are just the real values given by equations (14) through (16). Because the measured atomic number, neutron number and atomic mass number for subactinide nuclei in an external field are given by equations (14) through (16) it follows that these quantities are nondenumerable (non-integer) for the subactinides. It will be shown further on in this paper that this is not the case for actinide nuclei in an external field where now the measured atomic number, neutron number and atomic mass number are denumerable (integers).

The conventional liquid drop model of the nuclear binding energy is well described in the literature.<sup>28-36</sup> In the presence of an electromagnetic field the complex number binding energy of an atomic nucleus is written as<sup>26</sup>

$$\begin{aligned} \bar{B} &= \bar{E}_v - \bar{E}_s - \bar{E}_c - \bar{E}_{\text{sym}} + \bar{E}_{\text{pair}} + \bar{E}_{\text{shell}} \\ &= \bar{\alpha}\bar{a} - \bar{\gamma}\bar{a}^{2/3} - \bar{\delta}\bar{z}^2/\bar{a}^{1/3} - \bar{\beta}(\bar{n} - \bar{z})^2/\bar{a} + \bar{E}_{\text{pair}} + \bar{E}_{\text{shell}} \end{aligned} \quad (20)$$

where for subactinide nuclei  $\bar{z}$ ,  $\bar{n}$  and  $\bar{a}$  are given by equations (8) through (10) respectively,  $\bar{E}_v$ ,  $\bar{E}_s$ ,  $\bar{E}_c$ ,  $\bar{E}_{\text{sym}}$ ,  $\bar{E}_{\text{pair}}$  and  $\bar{E}_{\text{shell}}$  = complex number volume, surface, Coulomb, symmetry, pairing and shell energies respectively, and where  $\bar{\alpha}$ ,  $\bar{\gamma}$ ,  $\bar{\delta}$  and  $\bar{\beta}$  = complex number volume, surface, Coulomb and symmetry energy coefficients respectively. The mass formula coefficients are represented as

$$\bar{\alpha} = \alpha \exp(j\theta_\alpha) \quad \bar{\gamma} = \gamma \exp(j\theta_\gamma) \quad (21)$$

$$\bar{\delta} = \delta \exp(j\theta_\delta) \quad \bar{\beta} = \beta \exp(j\theta_\beta) \quad (22)$$

$$\begin{aligned} \bar{E}_{\text{pair}} &= E_{\text{pair}} \exp(j\theta_{E_{\text{pair}}}) & \bar{E}_{\text{shell}} &= E_{\text{shell}} \exp(j\theta_{E_{\text{shell}}}) \\ &= P\bar{\rho}/\bar{a}^{3/4} \end{aligned} \quad (23)$$

where

$$\bar{\rho} = \rho \exp(j\theta_\rho) \quad \rho \sim 34 \text{ MeV} \quad (24)$$

$$P = \begin{cases} 1 & Z \text{ even, } A \text{ even} \\ 0 & A \text{ odd} \\ -1 & Z \text{ odd, } A \text{ even} \end{cases} \quad (25)$$

and where for subactinide nuclei equation (12) and (23) give

$$E_{\text{pair}} = P\rho A^{-3/4} \cos^{-3/4}\theta_a \quad \theta_{\text{Epair}} = \theta_\rho - 3/4\theta_a \quad (26)$$

and where the shell energy has a more complicated variation. For the subactinide elements equations (8) through (10) and (20) through (26) give<sup>26</sup>

$$\begin{aligned} \bar{B} = & \alpha A \cos \theta_a \exp[j(\theta_\alpha + \theta_a)] - \gamma A^{2/3} \cos^{2/3}\theta_a \exp[j(\theta_\gamma + 2/3\theta_a)] \quad (27) \\ & - \delta Z^2 A^{-1/3} \cos^2\theta_z \cos^{-1/3}\theta_a \exp[j(\theta_\delta + 2\theta_z - 1/3\theta_a)] \\ & - \beta \xi^2 A^{-1} \cos^{-1}\theta_a \exp[j(\theta_\beta + 2\theta_\xi - \theta_a)] \\ & + P\rho A^{-3/4} \cos^{-3/4}\theta_a \exp[j(\theta_\rho - 3/4\theta_a)] \\ & + E_{\text{shell}} \exp(j\theta_{\text{Eshell}}) \end{aligned}$$

where the complex number neutron excess is given by<sup>26</sup>

$$\bar{\xi} = \xi \exp(j\theta_\xi) = \bar{n} - \bar{z} \quad (28)$$

The measured binding energy is given by the real part of equation (27) so that<sup>26</sup>

$$B_m = \alpha_m A - \gamma_m A^{2/3} - \delta_m Z^2 A^{-1/3} - \beta'_m \xi^2 A^{-1} + E_{\text{pair}}^m + E_{\text{shell}}^m \quad (29)$$

where for subactinide nuclei

$$\alpha_m = \alpha \cos \theta_a \cos(\theta_\alpha + \theta_a) \quad (30)$$

$$\gamma_m = \gamma \cos^{2/3}\theta_a \cos(\theta_\gamma + 2/3\theta_a) \quad (31)$$

$$\delta_m = \delta \cos^2\theta_z \cos^{-1/3}\theta_a \cos(\theta_\delta + 2\theta_z - 1/3\theta_a) \quad (32)$$

$$\beta'_m = \beta \cos^{-1}\theta_a \cos(\theta_\beta + 2\theta_\xi - \theta_a) \quad (33)$$

$$E_{\text{pair}}^m = P\rho A^{-3/4} \cos^{-3/4}\theta_a \cos(\theta_\rho - 3/4\theta_a) \quad (34)$$

The calculation of the  $\gamma$  ray flux that is required to catalyze thermal neutron induced fission in the subactinide elements using thermal neutrons has already appeared in the literature.<sup>26</sup> This calculation is based on the determination of the fission angle  $\theta_z^F$  of the atomic number which proceeds as follows. The complex number generalization of the Bohr-Wheeler fission instability condition is given by<sup>26</sup>

$$\bar{z}^2/\bar{a} = \kappa\bar{\gamma}/\bar{\delta} \quad (35)$$

or equivalently the two scalar fission stability boundary conditions are

$$z^2/a = \kappa\gamma/\delta \quad (36)$$

$$\theta_a = 2\theta_z - \theta_\gamma + \theta_\delta \quad (37)$$

Combining equation (12) with equations (36) and (37) gives the fission instability condition for subactinide nuclei located in an external field as<sup>26</sup>

$$\begin{aligned} z^2/A &= \kappa\gamma/\delta \cos \theta_a \cos^{-2} \theta_z \\ &= \kappa\gamma/\delta \cos(2\theta_z + \theta_\delta - \theta_\gamma) \cos^{-2} \theta_z \end{aligned} \quad (38)$$

Equation (38) can be solved to determine the value of  $\theta_z^F$  required to fission subactinide nuclei with thermal neutrons and the result is<sup>26</sup>

$$\tan \theta_z^F = \tan(\theta_\gamma - \theta_\delta) \pm \sec(\theta_\gamma - \theta_\delta) [1 - \chi \cos(\theta_\gamma - \theta_\delta)]^{1/2} \quad (39)$$

where the fissility parameter  $\chi$  is given by<sup>26</sup>

$$\chi = (z^2/A)(\kappa\gamma/\delta)^{-1} \quad (40)$$

and is restricted to the range  $0 \leq \chi \leq \sec(\theta_\gamma - \theta_\delta)$  by equation (39) and describes most subactinide nuclei except for some exotic proton rich subactinide nuclei for which  $\chi$  falls outside of this range. The value of the fission angle  $\theta_a^F$  of the atomic mass number that is required for the fission of subactinide nuclei by thermal neutrons is obtained from equation (37), as<sup>26</sup>

$$\theta_a^F = 2\theta_z^F - \theta_\gamma + \theta_\delta \quad (40A)$$

Figure 1 gives  $\theta_z^F$  and Figure 2 gives  $\theta_a^F$  for subactinide nuclei having  $0 \leq \chi \leq \sec(\theta_\gamma - \theta_\delta)$  and using the values of the angles  $\theta_\gamma = 0.4r$  and  $\theta_\delta = 0.1r$  whose values are selected for heuristic purposes. On the other hand, the assumption that  $\theta_\gamma = \theta_\delta$  allows equations (39) and (40A) to be written as<sup>26</sup>

$$\tan \theta_z^F \sim \pm (1 - \chi)^{1/2} \quad \theta_a^F \sim 2\theta_z^F \quad \theta_n^F \sim 3\theta_z^F \quad (41)$$

which are valid for  $0 \leq \chi \leq 1$ . Figure 3 gives  $\theta_z^F$  and Figure 4 gives  $\theta_a^F$  for subactinide nuclei under the approximation that  $\theta_\gamma = \theta_\delta$ .

The magnetic induction field of the  $\gamma$  rays that are required for the catalysis of thermal neutron induced fission of subactinide nuclei is then given by<sup>26</sup>

$$\begin{aligned} B_\gamma^F &= K_{\theta z}^{B\gamma} \tan \theta_z^F \\ &= K_{\theta z}^{B\gamma} (1 - \chi)^{1/2} \end{aligned} \quad (42)$$

where  $K_{\theta z}^{B\gamma}$  = dynamic magnetic stiffness modulus. Values of  $B_\gamma^F$  for various sub-

actinide nuclei have been tabulated in the literature.<sup>26</sup> This review of the theory of  $\gamma$  ray catalyzed fission of subactinide nuclei by thermal neutrons is sufficient to lead into the problem of quaternary fission of  $\gamma$  ray cooled actinide nuclei which is treated in the following sections.

This paper considers the thermal neutron induced quaternary fission of  $\gamma$  ray cooled actinide nuclei. It will be shown that the atomic number, neutron number and atomic mass number of the actinide elements in an external field must be represented by special kinds of complex numbers in internal space such that the measured atomic number, neutron number and atomic mass number are integers. These representations suggest that it is possible to suppress binary fission by thermal neutrons in the actinides by applying a  $\gamma$  ray field having the proper frequency and intensity. The result will be an enhanced quaternary fission rate in the actinides with fission products that are relatively low level radionuclides. Therefore clean fission of the actinide elements using thermal neutrons is possible by utilizing the cooling effects of a properly chosen  $\gamma$  ray bath. An outline of this paper is as follows: Section 2 gives the special forms of the complex atomic number, neutron number and atomic mass number for the actinide elements in an external field, Section 3 investigates the theory of the radioactive decay of actinide nuclei in an external electromagnetic field, Section 4 develops a liquid drop type of nuclear mass formula for actinide nuclei in an external field, Section 5 presents a theory of the suppression of thermal neutron induced binary fission of fissile actinide nuclei by an external field, Section 6 introduces the concept of thermal neutron induced quaternary fission in  $\gamma$  ray cooled actinide nuclei and gives the required  $\gamma$  ray photon energy, number density and flux density for binary fission suppression, and finally Section 7 gives the final state energy conditions for the binary fission of actinide nuclei in a low intensity  $\gamma$  ray field.

**2. ACTINIDE NUCLEI WITH BROKEN INTERNAL SYMMETRIES.** This section considers the broken internal symmetries of the atomic number, neutron number and atomic mass number of actinide nuclei that are subject to an external electromagnetic or gravitational field.

**A. Complex Atomic Number, Neutron Number and Atomic Mass Number for the Actinide Elements in an External Field.**

For those actinide nuclei which exhibit spontaneous or thermal neutron induced binary fission it is assumed that the component nucleons are in a measurably denumerable state so that the measured atomic number, neutron number and atomic mass number are integers even in the presence of an external field. This suggests that in the presence of an external field the atomic number, neutron number and atomic mass number of actinide nuclei are complex numbers of the form

$$\bar{z} = z \exp(j\theta_z) = Z \sec \theta_z \exp(j\theta_z) \quad (43)$$

$$\bar{n} = n \exp(j\theta_n) = N \sec \theta_n \exp(j\theta_n) \quad (44)$$

$$\bar{a} = a \exp(j\theta_a) = A \sec \theta_a \exp(j\theta_a) \quad (45)$$

and therefore for the actinides

$$z = Z \sec \theta_z \quad n = N \sec \theta_n \quad a = A \sec \theta_a \quad (46)$$

It is convenient to introduce the following terms

$$\bar{l}_z = \sec \theta_z \exp(j\theta_z) \quad (47)$$

$$\bar{l}_n = \sec \theta_n \exp(j\theta_n) \quad (48)$$

$$\bar{l}_a = \sec \theta_a \exp(j\theta_a) \quad (49)$$

Then equations (43) through (45) can be written as

$$\bar{z} = Z\bar{l}_z \quad \bar{n} = N\bar{l}_n \quad \bar{a} = A\bar{l}_a \quad (50)$$

The real and imaginary parts of equations (43) through (45) are written as

$$z_R = Z \quad n_R = N \quad a_R = A \quad (51)$$

$$z_I = Z \tan \theta_z \quad n_I = N \tan \theta_n \quad a_I = A \tan \theta_a \quad (52)$$

The measured values of the atomic number, neutron number and atomic mass number are just the corresponding real values so that

$$z_m = Z \quad n_m = N \quad a_m = A \quad (53)$$

and therefore the measured values of the atomic number, neutron number and atomic mass number for the actinides with  $\chi \geq 1$  in an external field are denumerable (integers). This is not the case for subactinide nuclei with  $\chi \leq 1$  in an external field as seen in equations (14) through (16).

The law of addition for the complex atomic number, complex neutron number and complex atomic mass number is given for both subactinide and actinide elements in an external field as<sup>26</sup>

$$W + \bar{a} = \bar{z} + \bar{n} \quad (54)$$

subject to the universal law of baryon number conservation

$$A = Z + N \quad (55)$$

The known quantities in equation (54) are taken to be  $Z$ ,  $\theta_z$ ,  $N$  and  $\theta_n$ . The value of  $A$  is determined from equation (55) while the unknown quantities  $W$  and  $\theta_a$  are obtained by taking the real and imaginary parts of equation (54) and using equations (51) and (52) with the result that for actinide nuclei



$$W + A = Z + N \quad (56)$$

$$A \tan \theta_a = Z \tan \theta_z + N \tan \theta_n \quad (57)$$

Therefore for actinide nuclei in an external field

$$W = 0 \quad (58)$$

$$\begin{aligned} \tan \theta_a &= Z/A \tan \theta_z + N/A \tan \theta_n \\ &= \tan \theta_n + Z/A(\tan \theta_z - \tan \theta_n) \end{aligned} \quad (59)$$

Equation (59) can also be written as

$$\tan \theta_n = (A \tan \theta_a - Z \tan \theta_z)/(A - Z) \quad (60)$$

$$\tan \theta_z = (A \tan \theta_a - N \tan \theta_n)/Z \quad (61)$$

which are valid for actinide nuclei. Combining equations (54) and (58) gives

$$\bar{a} = \bar{z} + \bar{n} \quad (62)$$

which is valid only for actinide nuclei because in this case the measured nucleon numbers are denumerable as shown in equation (51), so that  $W = 0$  for actinide nuclei as shown in equation (58). If  $\theta_z \sim \theta_n$  then equation (59) shows that the following condition is valid for the actinides

$$\theta_a \sim \theta_z \sim \theta_n \quad (63)$$

If  $\theta_a \sim 2\theta_z$  for actinide nuclei then the relationship between  $\theta_z$  and  $\theta_n$  is obtained from equation (59) to be

$$\tan(2\theta_z) - Z/A \tan \theta_z = N/A \tan \theta_n \quad (64)$$

The expressions for  $W$  and  $\theta_a$  for subactinide nuclei already appear in the literature.<sup>26</sup> For subactinide nuclei the measured nucleon numbers are nondenumerable as shown in equations (14) through (15) and  $W \neq 0$ .<sup>26</sup>

#### B. Time Variation of the Complex Atomic Number, Neutron Number and Atomic Mass Number for Actinide Nuclei.

This section considers the time variation of  $\bar{z}$ ,  $\bar{n}$  and  $\bar{a}$  for actinide nuclei. Two special cases are considered: the first is associated with changes in  $Z$ ,  $N$  and  $A$  due to nuclear fission and other radioactive decays in a constant external field, and the second is associated with the variations of  $\theta_z$ ,  $\theta_n$  and  $\theta_a$  that are related to a time varying external field for fixed values of  $Z$ ,  $N$  and  $A$ . The general case of the time variation of  $\bar{z}$ ,  $\bar{n}$  and  $\bar{a}$  is obtained from equations (43) through (45) as<sup>26</sup>

$$d\bar{z}/d\bar{t} = \cos \beta_{tt} (dz/dt + jz d\theta_z/dt) \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (65)$$

$$d\bar{n}/d\bar{t} = \cos \beta_{tt} (dn/dt + jn d\theta_n/dt) \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (66)$$

$$d\bar{a}/d\bar{t} = \cos \beta_{tt} (da/dt + jad\theta_a/dt) \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (67)$$

where  $\beta_{tt}$  is given by<sup>27</sup>

$$\tan \beta_{tt} = t \partial \theta_t / \partial t \quad (68)$$

The internal phase angles  $\theta_z$ ,  $\theta_n$  and  $\theta_a$  are functions of the external field, such as an electromagnetic field, so that

$$\theta_z = \theta_z(Z, H) \quad \theta_n = \theta_n(N, H) \quad \theta_a = \theta_a(A, H) \quad (69)$$

where  $H$  = amplitude of the magnetic field component of an electromagnetic field.

The time variation of the internal phase angles are written as

$$d\theta_z/dt = \partial \theta_z / \partial Z dZ/dt + \partial \theta_z / \partial H dH/dt \quad (70)$$

$$d\theta_n/dt = \partial \theta_n / \partial N dN/dt + \partial \theta_n / \partial H dH/dt \quad (71)$$

$$d\theta_a/dt = \partial \theta_a / \partial A dA/dt + \partial \theta_a / \partial H dH/dt \quad (72)$$

Combining equations (46) with equations (65) through (72) gives for the actinides

$$d\bar{z}/d\bar{t} = \cos \beta_{tt} (\bar{B}'_z dZ/dt + \bar{C}'_z dH/dt) \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (73)$$

$$d\bar{n}/d\bar{t} = \cos \beta_{tt} (\bar{B}'_n dN/dt + \bar{C}'_n dH/dt) \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (74)$$

$$d\bar{a}/d\bar{t} = \cos \beta_{tt} (\bar{B}'_a dA/dt + \bar{C}'_a dH/dt) \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (75)$$

where

$$\bar{B}'_z = \sec \theta_z [1 + (\tan \theta_z + j)Z \partial \theta_z / \partial Z] \quad (76)$$

$$\bar{B}'_n = \sec \theta_n [1 + (\tan \theta_n + j)N \partial \theta_n / \partial N] \quad (77)$$

$$\bar{B}'_a = \sec \theta_a [1 + (\tan \theta_a + j)A \partial \theta_a / \partial A] \quad (78)$$

$$\bar{C}'_z = \sec \theta_z (\tan \theta_z + j)Z \partial \theta_z / \partial H \quad (79)$$

$$\bar{C}'_n = \sec \theta_n (\tan \theta_n + j)N \partial \theta_n / \partial H \quad (80)$$

$$\bar{C}'_a = \sec \theta_a (\tan \theta_a + j)A \partial \theta_a / \partial H \quad (81)$$

where

$$\tan \theta_z + j = \sec \theta_z \exp[j(\pi/2 - \theta_z)] \quad (82)$$

$$\tan \theta_n + j = \sec \theta_n \exp[j(\pi/2 - \theta_n)] \quad (83)$$

$$\tan \theta_a + j = \sec \theta_a \exp[j(\pi/2 - \theta_a)] \quad (84)$$

Two special cases need to be considered.

Case a. Actinide Nuclei in a Constant Magnetic Field.

This case corresponds to nuclear reactions such as the fission of the actinides in a constant magnetic field and equations (73) through (75) become

$$(d\bar{z}/d\bar{t})_H = \cos \beta_{tt} \bar{B}'_z dZ/dt \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (85)$$

$$(d\bar{n}/d\bar{t})_H = \cos \beta_{tt} \bar{B}'_n dN/dt \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (86)$$

$$(d\bar{a}/d\bar{t})_H = \cos \beta_{tt} \bar{B}'_a dA/dt \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (87)$$

For this case the actual size of the nucleus is changing.

Case b. Actinides with Constant Z, N and A.

This case corresponds to changes in the internal phase angles  $\theta_z$ ,  $\theta_n$  and  $\theta_a$  in a time varying electromagnetic field. Then equations (73) through (75) give

$$(d\bar{z}/d\bar{t})_Z = \cos \beta_{tt} \bar{C}'_z dH/dt \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (88)$$

$$(d\bar{n}/d\bar{t})_N = \cos \beta_{tt} \bar{C}'_n dH/dt \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (89)$$

$$(d\bar{a}/d\bar{t})_A = \cos \beta_{tt} \bar{C}'_a dH/dt \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (90)$$

Case b corresponds to nuclei gaining or losing energy by internal space rotations.

The time derivatives of  $\bar{z}$ ,  $\bar{n}$  and  $\bar{a}$  given in equations (73) through (75) can also be written in a more general form involving only one exponential term by noting that equations (43) through (45) give

$$\begin{aligned} d\bar{z}/d\bar{t} &= \cos \beta_{tt} \sec \beta_{zz} dz/dt \exp(j\phi_{zt}) \\ &= \cos \beta_{tt} \csc \beta_{zz} z d\theta_z/dt \exp(j\phi_{zt}) \\ &= \sin \beta_{tt} \sec \beta_{zz} t^{-1} dz/d\theta_t \exp(j\phi_{zt}) \\ &= \sin \beta_{tt} \csc \beta_{zz} z/t d\theta_z/d\theta_t \exp(j\phi_{zt}) \end{aligned} \quad (91)$$

$$\begin{aligned}
d\bar{n}/d\bar{t} &= \cos \beta_{tt} \sec \beta_{nn} dn/dt \exp(j\phi_{nt}) \\
&= \cos \beta_{tt} \csc \beta_{nn} n d\theta_n/dt \exp(j\phi_{nt}) \\
&= \sin \beta_{tt} \sec \beta_{nn} t^{-1} dn/d\theta_t \exp(j\phi_{nt}) \\
&= \sin \beta_{tt} \csc \beta_{nn} n/t d\theta_n/d\theta_t \exp(j\phi_{nt})
\end{aligned} \tag{92}$$

$$\begin{aligned}
d\bar{a}/d\bar{t} &= \cos \beta_{tt} \sec \beta_{aa} da/dt \exp(j\phi_{at}) \\
&= \cos \beta_{tt} \csc \beta_{aa} a d\theta_a/dt \exp(j\phi_{at}) \\
&= \sin \beta_{tt} \sec \beta_{aa} t^{-1} da/d\theta_t \exp(j\phi_{at}) \\
&= \sin \beta_{tt} \csc \beta_{aa} a/t d\theta_a/d\theta_t \exp(j\phi_{at})
\end{aligned} \tag{93}$$

where  $\beta_{tt}$  is given by equation (68), and where

$$\tan \beta_{zz} = z \partial \theta_z / \partial z \quad \tan \beta_{nn} = n \partial \theta_n / \partial n \quad \tan \beta_{aa} = a \partial \theta_a / \partial a \tag{94}$$

$$\phi_{zt} = \theta_z + \beta_{zz} - \theta_t - \beta_{tt} \tag{95}$$

$$\phi_{nt} = \theta_n + \beta_{nn} - \theta_t - \beta_{tt} \tag{96}$$

$$\phi_{at} = \theta_a + \beta_{aa} - \theta_t - \beta_{tt} \tag{97}$$

The derivatives  $dz/dt$ ,  $dn/dt$  and  $da/dt$  are evaluated from equation (46) for actinide nuclei as

$$\begin{aligned}
dz/dt &= dZ/dt \sec \theta_z + Z \sec \theta_z \tan \theta_z d\theta_z/dt \\
&= \sec \theta_z [dZ/dt(1 + \tan \theta_z Z \partial \theta_z / \partial Z) + Z \tan \theta_z \partial \theta_z / \partial H dH/dt]
\end{aligned} \tag{98}$$

$$\begin{aligned}
dn/dt &= dN/dt \sec \theta_n + N \sec \theta_n \tan \theta_n d\theta_n/dt \\
&= \sec \theta_n [dN/dt(1 + \tan \theta_n N \partial \theta_n / \partial N) + N \tan \theta_n \partial \theta_n / \partial H dH/dt]
\end{aligned} \tag{99}$$

$$\begin{aligned}
da/dt &= dA/dt \sec \theta_a + A \sec \theta_a \tan \theta_a d\theta_a/dt \\
&= \sec \theta_a [dA/dt(1 + \tan \theta_a A \partial \theta_a / \partial A) + A \tan \theta_a \partial \theta_a / \partial H dH/dt]
\end{aligned} \tag{100}$$

It is assumed that  $\partial \theta_z / \partial t = 0$ ,  $\partial \theta_n / \partial t = 0$  and  $\partial \theta_a / \partial t = 0$ . The derivatives in equations (98) through (100) are used in equations (91) through (93).

Combining equations (46), (70) through (72), (91) through (93), and (98) through (100) gives for actinide nuclei

$$\begin{aligned} d\bar{z}/d\bar{t} &= J'_z \cos \beta_{tt} \sec \beta_{zz} \sec \theta_z \exp(j\phi_{zt}) \\ &= I_z \cos \beta_{tt} \csc \beta_{zz} \sec \theta_z \exp(j\phi_{zt}) \\ &= (I_z^2 + J_z'^2)^{1/2} \cos \beta_{tt} \sec \theta_z \exp(j\phi_{zt}) \end{aligned} \quad (101)$$

$$\begin{aligned} d\bar{n}/d\bar{t} &= J'_n \cos \beta_{tt} \sec \beta_{nn} \sec \theta_n \exp(j\phi_{nt}) \\ &= I_n \cos \beta_{tt} \csc \beta_{nn} \sec \theta_n \exp(j\phi_{nt}) \\ &= (I_n^2 + J_n'^2)^{1/2} \cos \beta_{tt} \sec \theta_n \exp(j\phi_{nt}) \end{aligned} \quad (102)$$

$$\begin{aligned} d\bar{a}/d\bar{t} &= J'_a \cos \beta_{tt} \sec \beta_{aa} \sec \theta_a \exp(j\phi_{at}) \\ &= I_a \cos \beta_{tt} \csc \beta_{aa} \sec \theta_a \exp(j\phi_{at}) \\ &= (I_a^2 + J_a'^2)^{1/2} \cos \beta_{tt} \sec \theta_a \exp(j\phi_{at}) \end{aligned} \quad (103)$$

where  $\phi_{zt}$ ,  $\phi_{nt}$  and  $\phi_{at}$  are given by equations (95) through (97) and where for actinide nuclei

$$\tan \beta_{zz} = I_z/J'_z \quad \tan \beta_{nn} = I_n/J'_n \quad \tan \beta_{aa} = I_a/J'_a \quad (104)$$

$$I_z = \tan \alpha_{ZZ}^H dZ/dt + Z/H \tan \alpha_{HH}^Z dH/dt \quad (105)$$

$$I_n = \tan \alpha_{NN}^H dN/dt + N/H \tan \alpha_{HH}^N dH/dt \quad (106)$$

$$I_a = \tan \alpha_{AA}^H dA/dt + A/H \tan \alpha_{HH}^A dH/dt \quad (107)$$

$$J'_z = (1 + \tan \theta_z \tan \alpha_{ZZ}^H) dZ/dt + Z/H \tan \theta_z \tan \alpha_{HH}^Z dH/dt \quad (108)$$

$$J'_n = (1 + \tan \theta_n \tan \alpha_{NN}^H) dN/dt + N/H \tan \theta_n \tan \alpha_{HH}^N dH/dt \quad (109)$$

$$J'_a = (1 + \tan \theta_a \tan \alpha_{AA}^H) dA/dt + A/H \tan \theta_a \tan \alpha_{HH}^A dH/dt \quad (110)$$

and where

$$\tan \alpha_{ZZ}^H = Z \partial \theta_z / \partial Z \quad \tan \alpha_{NN}^H = N \partial \theta_n / \partial N \quad \tan \alpha_{AA}^H = A \partial \theta_a / \partial A \quad (111)$$

$$\tan \alpha_{HH}^Z = H \partial \theta_z / \partial H \quad \tan \alpha_{HH}^N = H \partial \theta_n / \partial H \quad \tan \alpha_{HH}^A = H \partial \theta_a / \partial H \quad (112)$$

Equations (104) through (110) are valid for actinide nuclei in an external electromagnetic field.

Two special cases are of interest for application to nuclear reactions and transmutations of the actinides.

Case a. External Radioactive Decay in a Constant Electromagnetic Field.

For this case it follows from equations (105) through (110) that for the actinides

$$I_z^H = \tan \alpha_{ZZ}^H dZ/dt \quad (113)$$

$$I_n^H = \tan \alpha_{NN}^H dN/dt \quad (114)$$

$$I_a^H = \tan \alpha_{AA}^H dA/dt \quad (115)$$

$$J_z^{H'} = (1 + \tan \theta_z \tan \alpha_{ZZ}^H) dZ/dt \quad (116)$$

$$J_n^{H'} = (1 + \tan \theta_n \tan \alpha_{NN}^H) dN/dt \quad (117)$$

$$J_a^{H'} = (1 + \tan \theta_a \tan \alpha_{AA}^H) dA/dt \quad (118)$$

$$\tan \beta_{zz}^H = I_z^H / J_z^{H'} = \tan \alpha_{ZZ}^H (1 + \tan \theta_z \tan \alpha_{ZZ}^H)^{-1} \quad (119)$$

$$\tan \beta_{nn}^H = I_n^H / J_n^{H'} = \tan \alpha_{NN}^H (1 + \tan \theta_n \tan \alpha_{NN}^H)^{-1} \quad (120)$$

$$\tan \beta_{aa}^H = I_a^H / J_a^{H'} = \tan \alpha_{AA}^H (1 + \tan \theta_a \tan \alpha_{AA}^H)^{-1} \quad (121)$$

$$\alpha_{HH}^Z = \pi/2 \quad \alpha_{HH}^N = \pi/2 \quad \alpha_{HH}^A = \pi/2 \quad (122)$$

From equations (85) through (87) and (101) through (122) it follows for constant H that for the actinides

$$\begin{aligned} (d\bar{z}/d\bar{t})_H &= J_z^{H'} \cos \beta_{tt} \sec \beta_{zz}^H \sec \theta_z \exp(j\phi_{zt}^H) \\ &= I_z^H \cos \beta_{tt} \csc \beta_{zz}^H \sec \theta_z \exp(j\phi_{zt}^H) \\ &= [(I_z^H)^2 + (J_z^{H'})^2]^{1/2} \cos \beta_{tt} \sec \theta_z \exp(j\phi_{zt}^H) \end{aligned} \quad (123)$$

$$\begin{aligned}
(d\bar{n}/d\bar{t})_H &= J_n^{H'} \cos \beta_{tt} \sec \beta_{nn}^H \sec \theta_n \sec(j\phi_{nt}^H) \\
&= I_n^H \cos \beta_{tt} \csc \beta_{nn}^H \sec \theta_n \exp(j\phi_{nt}^H) \\
&= [(I_n^H)^2 + (J_n^{H'})^2]^{1/2} \cos \beta_{tt} \sec \theta_n \exp(j\phi_{nt}^H)
\end{aligned} \tag{124}$$

$$\begin{aligned}
(d\bar{a}/d\bar{t})_H &= J_a^{H'} \cos \beta_{tt} \sec \beta_{aa}^H \sec \theta_a \exp(j\phi_{at}^H) \\
&= I_a^H \cos \beta_{tt} \csc \beta_{aa}^H \sec \theta_a \exp(j\phi_{at}^H) \\
&= [(I_a^H)^2 + (J_a^{H'})^2]^{1/2} \cos \beta_{tt} \sec \theta_a \exp(j\phi_{at}^H)
\end{aligned} \tag{125}$$

where

$$\phi_{zt}^H = \theta_z + \beta_{zz}^H - \theta_t - \beta_{tt} \tag{126}$$

$$\phi_{nt}^H = \theta_n + \beta_{nn}^H - \theta_t - \beta_{tt} \tag{127}$$

$$\phi_{at}^H = \theta_a + \beta_{aa}^H - \theta_t - \beta_{tt} \tag{128}$$

where  $\beta_{zz}^H$ ,  $\beta_{nn}^H$  and  $\beta_{aa}^H$  are given by equations (119) through (121). It is easy to see that equations (85) through (87) are equivalent to equations (123) through (125).

Case b. Internal Phase Radioactive Decay of the Actinides  
due to a Time Dependent Electromagnetic Field.

The case at hand corresponds to constant values of  $Z$ ,  $N$  and  $A$ . Equations (104) through (112) give for the actinides

$$\tan \beta_{zz}^Z = \cot \theta_z \quad \beta_{zz}^Z = \pi/2 - \theta_z \tag{129}$$

$$\tan \beta_{nn}^N = \cot \theta_n \quad \beta_{nn}^N = \pi/2 - \theta_n \tag{130}$$

$$\tan \beta_{aa}^A = \cot \theta_a \quad \beta_{aa}^A = \pi/2 - \theta_a \tag{131}$$

$$\csc \beta_{zz}^Z = \sec \theta_z \quad \csc \beta_{nn}^N = \sec \theta_n \quad \csc \beta_{aa}^A = \sec \theta_a \tag{132}$$

$$\sec \beta_{zz}^Z = \csc \theta_z \quad \sec \beta_{nn}^N = \csc \theta_n \quad \sec \beta_{aa}^A = \csc \theta_a \tag{133}$$

$$\alpha_{ZZ}^H = \pi/2 \quad \alpha_{NN}^H = \pi/2 \quad \alpha_{AA}^H = \pi/2 \tag{134}$$

$$I_z^Z = Z/H \tan \alpha_{HH}^Z dH/dt \quad (135)$$

$$I_n^N = N/H \tan \alpha_{HH}^N dH/dt \quad (136)$$

$$I_a^A = A/H \tan \alpha_{HH}^A dH/dt \quad (137)$$

$$J_z^{Z'} = Z/H \tan \theta_z \tan \alpha_{HH}^Z dH/dt \quad (138)$$

$$J_n^{N'} = N/H \tan \theta_n \tan \alpha_{HH}^N dH/dt \quad (139)$$

$$J_a^{A'} = A/H \tan \theta_a \tan \alpha_{HH}^A dH/dt \quad (140)$$

Combining equations (135) through (140) and using equation (112) gives the following results for constant Z, N and A

$$\begin{aligned} [(I_z^Z)^2 + (J_z^{Z'})^2]^{1/2} &= Z/H \tan \alpha_{HH}^Z \sec \theta_z dH/dt \\ &= Z \partial \theta_z / \partial H \sec \theta_z dH/dt \end{aligned} \quad (141)$$

$$\begin{aligned} [(I_n^N)^2 + (J_n^{N'})^2]^{1/2} &= N/H \tan \alpha_{HH}^N \sec \theta_n dH/dt \\ &= N \partial \theta_n / \partial H \sec \theta_n dH/dt \end{aligned} \quad (142)$$

$$\begin{aligned} [(I_a^A)^2 + (J_a^{A'})^2]^{1/2} &= A/H \tan \alpha_{HH}^A \sec \theta_a dH/dt \\ &= A \partial \theta_a / \partial H \sec \theta_a dH/dt \end{aligned} \quad (143)$$

Combining equations (101) through (103) with equations (141) through (143) gives for the actinides with constant Z, N and A

$$(d\bar{z}/d\bar{t})_Z = Z \cos \beta_{tt} \sec^2 \theta_z \partial \theta_z / \partial H dH/dt \exp(j\phi_{zt}^Z) \quad (144)$$

$$(d\bar{n}/d\bar{t})_N = N \cos \beta_{tt} \sec^2 \theta_n \partial \theta_n / \partial H dH/dt \exp(j\phi_{nt}^N) \quad (145)$$

$$(d\bar{a}/d\bar{t})_A = A \cos \beta_{tt} \sec^2 \theta_a \partial \theta_a / \partial H dH/dt \exp(j\phi_{at}^A) \quad (146)$$

where equations (95) through (97) and equations (129) through (131) give

$$\begin{aligned} \phi_{zt}^Z &= \theta_z + \beta_{zz}^Z - \theta_t - \beta_{tt} \\ &= \pi/2 - \theta_t - \beta_{tt} \end{aligned} \quad (147)$$



$$\begin{aligned}\phi_{nt}^N &= \theta_n + \beta_{nn}^N - \theta_t - \beta_{tt} \\ &= \pi/2 - \theta_t - \beta_{tt}\end{aligned}\quad (148)$$

$$\begin{aligned}\phi_{at}^A &= \theta_a + \beta_{aa}^A - \theta_t - \beta_{tt} \\ &= \pi/2 - \theta_t - \beta_{tt}\end{aligned}\quad (149)$$

Therefore for the internal phase nuclear reactions of the actinides

$$\phi_{zt}^Z = \phi_{nt}^N = \phi_{at}^A = \pi/2 - \theta_t - \beta_{tt} \quad (150)$$

which is a result that is different from the analogous case for the subactinide elements.<sup>26</sup> Equations (144) through (146) can also be written as

$$(\partial \bar{z} / \partial H)_Z = jZ \sec^2 \theta_Z \partial \theta_Z / \partial H \quad (151)$$

$$(\partial \bar{n} / \partial H)_N = jN \sec^2 \theta_n \partial \theta_n / \partial H \quad (152)$$

$$(\partial \bar{a} / \partial H)_A = jA \sec^2 \theta_a \partial \theta_a / \partial H \quad (153)$$

which actually corresponds to

$$(d\bar{z})_Z = jZ \sec^2 \theta_Z d\theta_Z \quad (154)$$

$$(d\bar{n})_N = jN \sec^2 \theta_n d\theta_n \quad (155)$$

$$(d\bar{a})_A = jA \sec^2 \theta_a d\theta_a \quad (156)$$

which is easily derived from equations (43) through (45) and equations (82) through (84) for the internal phase rotations of the actinides.

3. RADIOACTIVE DECAY OF THE ACTINIDE NUCLEI IN AN EXTERNAL FIELD. This section investigates the radioactive decay of a collection of actinide nuclei in the presence of an external electromagnetic or gravitational field. Internal phase angles are induced into the number of nuclei, into the total number of constituent nucleons and into the atomic mass number of each nucleus. An addition law for actinide nuclei is developed that relates the complex total nucleon number, the complex number of atomic nuclei and the complex atomic mass number.

A. Addition Law for Actinide Nuclei with Complex Total Nucleon Number, Complex Number of Atomic Nuclei and Complex Atomic Mass Number.

In the presence of an external field such as gravity or an electromagnetic

field the particle number of any system of particles must be represented by complex numbers in an internal space.<sup>26</sup> Therefore for nuclei in an electromagnetic field the complex number of total nucleons (protons and neutrons within the nuclei), the complex number of atomic nuclei and the complex atomic mass number are written as<sup>26</sup>

$$\bar{N}_n = N_n \exp(j\theta_{Nn}) \quad (157)$$

$$\bar{N} = N \exp(j\theta_N) \quad (158)$$

$$\bar{a} = a \exp(j\theta_a) \quad (159)$$

where  $\bar{N}_n$ ,  $N_n$  and  $\theta_{Nn}$  = complex number value, magnitude and internal phase angle of the total number of nucleons situated within all of the atomic nuclei;  $\bar{N}$ ,  $N$  and  $\theta_N$  = complex number value, magnitude and internal phase angle of the number of atomic nuclei; and as before  $\bar{a}$ ,  $a$  and  $\theta_a$  = complex number value, magnitude and internal phase angle of the atomic mass number of each nucleus. For actinide nuclei it follows that in analogy to equations (43) through (46) the particle number magnitudes that appear in equations (157) through (159) are written as

$$N_n = \eta_n \sec \theta_{Nn} \quad (160)$$

$$N = \eta \sec \theta_N \quad (161)$$

$$a = A \sec \theta_a \quad (162)$$

where  $\eta_n$  = integer number of total number of nucleons within the atomic nuclei,  $\eta$  = integer number of atomic nuclei, and as before  $A$  = atomic mass number which is an integer. The integer numbers satisfy the following fundamental universally valid law of baryon number conservation

$$\eta_n = \eta A \quad (163)$$

The measured nucleon number, nuclei number and atomic mass number are given by the real parts of equations (157) through (159). Using equations (160) through (162) gives the result

$$\left. \begin{aligned} N_{mn} &= \eta_n \\ N_m &= \eta \\ a_m &= A \end{aligned} \right\} \quad (164)$$

which are integers for actinide nuclei. For actinide nuclei the measured nucleon number, nuclei number and atomic mass number are denumerable. This is not the case for subactinide nuclei.<sup>26</sup>

Following the example of the addition law given by equation (54) which is valid for a single nucleus, the addition law for the complex particle numbers is written as<sup>26</sup>

$$W' + \bar{N}_n = \eta \bar{a} \quad (165)$$

which is subject to the universal validity of equation (163). The component

equations of equation (165) for actinide nuclei are written with the help of equation (164) as

$$W' + \eta_n = \eta A \quad (166)$$

$$\eta_n \tan \theta_{Nn} = \eta A \tan \theta_a \quad (167)$$

which give the relations

$$W' = 0 \quad \theta_{Nn} = \theta_a \quad (168)$$

which are valid for actinide nuclei. Therefore for actinide nuclei

$$\bar{N}_n = \eta \bar{a} \quad N_n = \eta a \quad \theta_{Nn} = \theta_a \quad (169)$$

and equation (169) is equivalent to the integer relationship given by equation (163). This situation is much simpler than the case for subactinide nuclei where  $W' \neq 0$ .<sup>26</sup>

#### B. Radioactive Decay of Actinide Nuclei in the Presence of an Electromagnetic Field.

This section considers the radioactive decay of the actinides in a time varying external field such as electromagnetism or gravitation. The complex number generalization of the standard radioactive decay law for the heavy elements is written as<sup>26</sup>

$$d\bar{N}/d\bar{t} = -\bar{\lambda}\bar{N} \quad (170)$$

where  $\bar{N}$  = complex number of atomic nuclei and  $\bar{\lambda}$  = complex number radioactive decay constant which can be written as<sup>26</sup>

$$\bar{\lambda} = \lambda \exp(j\theta_\lambda) \quad (171)$$

From equations (158) and (161) it follows that  $\bar{N}$  can be written for actinide elements as

$$\begin{aligned} \bar{N} &= N \exp(j\theta_N) & N &= \eta \sec \theta_N \\ &= \eta \sec \theta_N \exp(j\theta_N) \end{aligned} \quad (172)$$

The time derivative of equation (172) is given by<sup>26</sup>

$$d\bar{N}/d\bar{t} = \cos \beta_{tt} \sec \beta_{NN} dN/dt \exp(j\phi_{Nt}) \quad (173)$$

$$= \cos \beta_{tt} \csc \beta_{NN} N d\theta_N/dt \exp(j\phi_{Nt}) \quad (174)$$

$$= \sin \beta_{tt} \sec \beta_{NN} t^{-1} dN/d\theta_t \exp(j\phi_{Nt}) \quad (175)$$

$$= \sin \beta_{tt} \csc \beta_{NN} N/t d\theta_N/d\theta_t \exp(j\phi_{Nt}) \quad (176)$$

where<sup>26</sup>

$$\tan \beta_{NN} = N \partial \theta_N / \partial N \quad (177)$$

$$\phi_{Nt} = \theta_N + \beta_{NN} - \theta_t - \beta_{tt} \quad (178)$$

Then the law of radioactive decay of elements given in equation (170) can be written in any of the following forms<sup>26</sup>

$$\cos \beta_{tt} \sec \beta_{NN} dN/dt = -\lambda N \quad (179)$$

$$\cos \beta_{tt} \csc \beta_{NN} d\theta_N/dt = -\lambda \quad (180)$$

$$\sin \beta_{tt} \sec \beta_{NN} t^{-1} dN/d\theta_t = -\lambda N \quad (181)$$

$$\sin \beta_{tt} \csc \beta_{NN} t^{-1} d\theta_N/d\theta_t = -\lambda \quad (182)$$

combined with the following phase angle relationship

$$\phi_{Nt} = \theta_\lambda + \theta_N \quad (183)$$

Combining equations (178) and (183) gives<sup>26</sup>

$$\beta_{NN} - \theta_t - \beta_{tt} = \theta_\lambda \quad (184)$$

The derivative  $d\theta_N/dt$  that appears in equation (174) is written as

$$d\theta_N/dt = \partial \theta_N / \partial \eta \, d\eta/dt + \partial \theta_N / \partial H \, dH/dt \quad (185)$$

The derivative  $dN/dt$  that appears in equation (173) is obtained from equation (172) to be for actinide nuclei

$$\begin{aligned} dN/dt &= d\eta/dt \sec \theta_N + \eta \sec \theta_N \tan \theta_N \, d\theta_N/dt \\ &= \sec \theta_N [d\eta/dt (1 + \tan \theta_N \, \eta \partial \theta_N / \partial \eta) + \eta \tan \theta_N \, \partial \theta_N / \partial H \, dH/dt] \end{aligned} \quad (186)$$

Equations (170), (171) and (173) through (185) are valid for the radioactive decay of all nuclei but equations (172) and (186) refer only to actinide nuclei.

Another form of the law of radioactive decay that applies only to the actinides can be obtained in analogy to equations (101) through (103) by noting that the time derivative that appears in the radioactivity law in equation (170) can be obtained from equations (173), (174), (185) and (186) as

$$d\bar{N}/d\bar{t} = \cos \beta_{tt} \sec \beta_{NN} \sec \theta_N J'_N \exp(j\phi_{Nt}) \quad (187)$$

$$= \cos \beta_{tt} \csc \beta_{NN} \sec \theta_N I_N \exp(j\phi_{Nt}) \quad (188)$$

$$= \cos \beta_{tt} \sec \theta_N (I_N^2 + J_N'^2)^{1/2} \exp(j\phi_{Nt}) \quad (189)$$

where

$$I_N = \tan \alpha_{\eta\eta}^H \, d\eta/dt + \eta/H \tan \alpha_{HH}^{\eta} \, dH/dt \quad (190)$$

$$J'_N = (1 + \tan \theta_N \tan \alpha_{\eta\eta}^H) d\eta/dt + \eta/H \tan \theta_N \tan \alpha_{HH}^{\eta} \, dH/dt \quad (191)$$

$$\tan \beta_{NN} = N \partial \theta_N / \partial N = I_N / J'_N \quad (192)$$

$$\tan \alpha_{\eta\eta}^H = \eta \partial \theta_N / \partial \eta \quad (193)$$

$$\tan \alpha_{HH}^{\eta} = H \partial \theta_N / \partial H \quad (194)$$

Combining equations (170), (172) and (189) gives for the actinides

$$\begin{aligned} \cos \beta_{tt} \sec \theta_N (I_N^2 + J_N'^2)^{1/2} \exp(j\phi_{Nt}) \\ = -\lambda \eta \sec \theta_N \exp[j(\theta_\lambda + \theta_N)] \end{aligned} \quad (195)$$

which gives the radioactive decay law as

$$\cos \beta_{tt} (I_N^2 + J_N'^2)^{1/2} = -\lambda \eta \quad (196)$$

along with the phase angle relationship given by equations (183) and (184). Equation (196) is equivalent to equations (179) and (180). Equation (196) for the radioactive decay of the actinides is similar to equation (255) of Reference 26 which describes the radioactive decay of subactinide nuclei. The difference is the + signs that occur in the expression for  $J'_N$  in equation (191) for actinide nuclei, and the corresponding negative signs that occur in the expression for  $J_N$  that describes the radioactive decay of subactinide nuclei as in equations (250) and (255) of Reference 26. The phase angle  $\theta_N$  may be negative so that in this case

$$\tan \theta_N = -\tan |\theta_N| \quad (197)$$

The angle  $\theta_N$  can be positive or negative.

The case of incoherent radioactive decay of the actinides can be regained by taking  $\theta_N = \text{constant}$ ,  $\theta_t = 0$  and  $\beta_{tt} = 0$  in which case equations (170), (173) and (186) become

$$\sec \theta_N \, d\eta/dt \exp(j\theta_N) = -\lambda \eta \sec \theta_N \exp[j(\theta_N + \theta_\lambda)] \quad (198)$$

which gives the conventional law of radioactive decay of elements as<sup>37,38</sup>

$$d\eta/dt = -\lambda \eta \quad \theta_\lambda = 0 \quad (199)$$

This case can also be obtained from equations (187) through (196) by recognizing that the case of incoherent radioactivity corresponds to

$$\beta_{NN}^H = 0 \quad \alpha_{\eta\eta}^H = 0 \quad \alpha_{HH}^\eta = 0 \quad (200)$$

$$J_N^H = d\eta/dt \quad I_N^H = 0 \quad (201)$$

for which case equation (196) reduces to equation (199).

The case of radioactive decay of the actinides in a static magnetic field is obtained from equations (190), (191) and (196) and is described by

$$I_N^H = \tan \alpha_{\eta\eta}^H d\eta/dt \quad (202)$$

$$J_N^H = (1 + \tan \theta_N \tan \alpha_{\eta\eta}^H) d\eta/dt \quad (203)$$

$$Y' \cos \beta_{tt} d\eta/dt = -\lambda\eta \quad (204)$$

where

$$Y' = [\tan^2 \alpha_{\eta\eta}^H + (1 + \tan \theta_N \tan \alpha_{\eta\eta}^H)^2]^{1/2} \quad (205)$$

Equation (204) can be rewritten as

$$d\eta/dt = -\lambda''\eta \quad (206)$$

where the effective decay constant for the actinides is given by

$$\lambda'' = \lambda/Y' \sec \beta_{tt} \quad (207)$$

For the actinides the effective radioactive decay constant is a decreasing function of the strength of the applied magnetic field if  $\partial\alpha_{\eta\eta}^H/\partial H > 0$ , and vice versa. The phase angle equation for the case of a static magnetic field is obtained by first realizing that for  $H = \text{constant}$  equation (192) becomes

$$\tan \beta_{NN}^H = I_N^H/J_N^H = \tan \alpha_{\eta\eta}^H (1 + \tan \theta_N \tan \alpha_{\eta\eta}^H)^{-1} \quad (208)$$

and equation (184) gives the phase angle condition as

$$\theta_\lambda = \beta_{NN}^H - \theta_t - \beta_{tt} \quad (209)$$

Equation (209) gives the internal phase angle of the radioactive decay constant for atomic nuclei in a constant magnetic field.

### C. Coherent Radioactive Decay of the Actinides.

This section considers the case where  $\eta = \text{constant}$  which corresponds to radioactive decays where the integer number of actinide nuclei remains constant

and only the internal phase angle  $\theta_N$  changes due to the presence of a time dependent external electromagnetic field. For  $\eta = \text{constant}$ , equations (190), (191) and (194) give for actinide nuclei

$$\begin{aligned} I_N^\eta &= \eta/H \tan \alpha_{HH}^\eta dH/dt \\ &= \eta d\theta_N/dt \end{aligned} \quad (210)$$

$$\begin{aligned} J_N^{\eta'} &= \eta/H \tan \theta_N \tan \alpha_{HH}^\eta dH/dt \\ &= \eta \tan \theta_N d\theta_N/dt \end{aligned} \quad (211)$$

and equation (192) gives

$$\tan \beta_{NN}^\eta = \cot \theta_N \quad \beta_{NN}^\eta = \pi/2 - \theta_N \quad (212)$$

Equations (196), (210), (211) and equations (184) and (212) give the internal phase radioactive decay equations for actinide nuclei as follows

$$\cos \beta_{tt} \sec \theta_N d\theta_N/dt = -\lambda \quad (213)$$

$$\begin{aligned} \theta_\lambda &= \beta_{NN}^\eta - \theta_t - \beta_{tt} \\ &= \pi/2 - \theta_N - \theta_t - \beta_{tt} \end{aligned} \quad (214)$$

where  $\beta_{tt}$  is given by equation (68). The internal phase radioactive decay equation (213) can also be obtained from equations (180) and (212). Equations (213) and (214) can also be obtained directly from equations (170) and (172) or from equations (189) through (191) by realizing that for actinide nuclei with  $\eta = \text{constant}$

$$\begin{aligned} (d\bar{N}/d\bar{t})_\eta &= \eta \cos \beta_{tt} \sec^2 \theta_N d\theta_N/dt \exp[j(\pi/2 - \theta_t - \beta_{tt})] \\ &= \bar{N} \cos \beta_{tt} \sec \theta_N d\theta_N/dt \exp[j(\pi/2 - \theta_N - \theta_t - \beta_{tt})] \end{aligned} \quad (215)$$

and

$$\begin{aligned} (d\bar{N})_\eta &= \eta \sec^2 \theta_N d\theta_N \exp(j\pi/2) \\ &= j\eta \sec^2 \theta_N d\theta_N \\ &= j\bar{N} \sec \theta_N d\theta_N \exp(-j\theta_N) \\ &= \bar{N} \sec \theta_N d\theta_N \exp[j(\pi/2 - \theta_N)] \end{aligned} \quad (216)$$

and where

$$|(d\bar{N})_\eta| = \eta \sec^2 \theta_N d\theta_N \quad (217)$$

Equation (216) gives the coherent ( $\eta = \text{constant}$ ) change of the complex number of actinide nuclei given by equation (172). Equations (213) and (214) are coupled simultaneous differential equations that determine  $\theta_N$  and  $\theta_t$  for actinide nuclei that are undergoing internal phase radioactive decay. For slowly changing values of  $\theta_t$  equation (214) is written as

$$\theta_\lambda = \pi/2 - \theta_N - \theta_t - t\partial\theta_t/\partial t \quad (218)$$

which is an approximate differential equation for  $\theta_N$  and  $\theta_t$ .

It has been shown that a solution to equation (213) is given by<sup>26</sup>

$$(\sec \theta_N + \tan \theta_N)/(\sec \theta_N^0 + \tan \theta_N^0) = \exp[-\lambda g(t)] \quad (219)$$

where

$$g(t) = \int_0^t \sec \beta_{tt} dt \quad (220)$$

where  $\theta_N = \theta_N^0$  for  $t = 0$ , and for  $t = \infty$  it follows that<sup>26</sup>

$$\theta_N^\infty = -\pi/2 \quad (221)$$

From equation (214) it follows that for  $t \rightarrow \infty$  and for actinide nuclei

$$\begin{aligned} \theta_\lambda &= \pi - \theta_t - \beta_{tt} \\ &\sim \pi - \theta_t - t\partial\theta_t/\partial t \end{aligned} \quad (222)$$

which is a differential equation for  $\theta_t$ . Therefore for the internal phase radioactive decay of actinide nuclei in the limit of  $t \rightarrow \infty$  equation (222) gives

$$\theta_t = \pi - \theta_\lambda + c/t \quad (223)$$

where  $c = \text{constant}$ . For  $t = \infty$  equation (223) gives

$$\theta_t^\infty = \pi - \theta_\lambda \quad (224)$$

for internal phase decays of actinide nuclei.

**4. NUCLEAR MASS FORMULA FOR ACTINIDE NUCLEI IN AN ELECTROMAGNETIC OR GRAVITATIONAL FIELD.** This section develops a broken symmetry form of the liquid drop type or nuclear mass formula that describes the nuclear binding energy of actinide nuclei. The nuclear mass formula can be used to study the suppression of thermal neutron induced binary fission and the enhancement of quaternary fission in the actinides by the presence of a  $\gamma$  ray field as described in Sections 5 and 6.

A. Complex Number Radius for Actinide Nuclei.



The simplest scalar expression for the nuclear radius is given by<sup>28-36</sup>

$$R = bA^{1/3} \quad b = 1.2 \text{ fm} \quad (225)$$

In an external electromagnetic or gravitational field the nuclear radius is represented by a complex number as follows<sup>26</sup>

$$\bar{R} = R \exp(j\theta_R) = \bar{b}\bar{a}^{1/3} \quad (226)$$

where the complex number constant  $\bar{b}$  is written as<sup>26</sup>

$$\bar{b} = b \exp(j\theta_b) \quad (227)$$

and  $\bar{a}$  = complex atomic mass number which for the actinides is given by equation (45). Combining equations (45) and (226) gives for the actinides

$$\bar{R} = bA^{1/3} \sec^{1/3}\theta_a \exp[j(\theta_b + \theta_a/3)] \quad (228)$$

$$R = bA^{1/3} \sec^{1/3}\theta_a \quad \theta_R = \theta_b + \theta_a/3 \quad (229)$$

The measured radius of an actinide nucleus is given by the real part of equation (228)

$$R_m = R \cos \theta_R = bA^{1/3} \sec^{1/3}\theta_a \cos(\theta_b + \theta_a/3) \quad (230)$$

which can be rewritten as

$$R_m = b'A^{1/3} \quad (231)$$

where for the actinides

$$b' = b \sec^{1/3}\theta_a \cos(\theta_b + \theta_a/3) \quad (232)$$

The effective radius constant  $b'$  is a slowly increasing function of the applied external field. Therefore in a weak external field  $b' \sim b = 1.2 \text{ fm}$ , while in a strong field  $b' > b$ .

#### B. Binding Energy of Actinide Nuclei Located in an Electromagnetic or Gravitational Field.

The standard scalar expression for the binding energy  $B$  of a nucleus  $(Z, A)$  is given by the liquid drop model as<sup>28-35</sup>

$$\begin{aligned} B &= E_v - E_s - E_c - E_{\text{sym}} + E_{\text{pair}} + E_{\text{shell}} \\ &= \alpha A - \gamma A^{2/3} - \delta Z^2/A^{1/3} - \beta(N - Z)^2/A + P\rho/A^{3/4} + E_{\text{shell}} \end{aligned} \quad (233)$$

where  $E_v$ ,  $E_s$ ,  $E_c$ ,  $E_{\text{sym}}$ ,  $E_{\text{pair}}$  and  $E_{\text{shell}}$  = volume, surface, Coulomb, symmetry, nuclear pairing and nuclear shell energies respectively, and where  $\alpha$ ,

$\gamma$  ,  $\delta$  and  $\beta$  = volume, surface, Coulomb and symmetry energy coefficients respectively, and where from equation (55) it follows that  $N - Z = A - 2Z$  . The pairing energy is given by<sup>28-35</sup>

$$E_{\text{pair}} = P\rho/A^{3/4} \quad (234)$$

where  $\rho$  = pairing energy coefficient  $\sim 34$  MeV and where  $P$  is given by equation (25). The average binding energy per nucleon  $\epsilon = B/A$  is written as

$$\begin{aligned} \epsilon &= \epsilon_v - \epsilon_s - \epsilon_c - \epsilon_{\text{sym}} + \epsilon_{\text{pair}} + \epsilon_{\text{shell}} \\ &= \alpha - \gamma/A^{1/3} - \delta Z^2/A^{4/3} - \beta[(N - Z)/A]^2 + P\rho/A^{7/4} + \epsilon_{\text{shell}} \end{aligned} \quad (235)$$

where  $\epsilon_v$  ,  $\epsilon_s$  ,  $\epsilon_c$  ,  $\epsilon_{\text{sym}}$  ,  $\epsilon_{\text{pair}}$  and  $\epsilon_{\text{shell}}$  = average volume energy per nucleon, average surface energy per nucleon, average Coulomb energy per nucleon, average symmetry energy per nucleon, average pairing energy per nucleon and the average shell energy per nucleon respectively. More complicated forms of the nuclear symmetry energy have been considered by including the effects of the nuclear bulk modulus.<sup>36</sup> However, in this paper only the simple Weizsäcker-Bethe form given in equation (235) is considered.

For an actinide nucleus in the presence of an electromagnetic or gravitational field the complex number nuclear binding energy is written as<sup>26</sup>

$$\begin{aligned} \bar{B} &= \bar{E}_v - \bar{E}_s - \bar{E}_c - \bar{E}_{\text{sym}} + \bar{E}_{\text{pair}} + \bar{E}_{\text{shell}} \\ &= \bar{\alpha}\bar{a} - \bar{\gamma}\bar{a}^{-2/3} - \bar{\delta}\bar{z}^2/\bar{a}^{1/3} - \bar{\beta}(\bar{n} - \bar{z})^2/\bar{a} + P\bar{\rho}/\bar{a}^{3/4} + \bar{E}_{\text{shell}} \end{aligned} \quad (236)$$

where  $\bar{z}$  ,  $\bar{n}$  and  $\bar{a}$  are given by equations (43) through (45) respectively,  $\bar{E}_v$  ,  $\bar{E}_s$  ,  $\bar{E}_c$  ,  $\bar{E}_{\text{sym}}$  ,  $\bar{E}_{\text{pair}}$  and  $\bar{E}_{\text{shell}}$  = complex number volume, surface, Coulomb, symmetry, pairing and shell energies respectively, and where  $\bar{\alpha}$  ,  $\bar{\gamma}$  ,  $\bar{\delta}$  and  $\bar{\beta}$  = complex number volume surface, Coulomb and symmetry energy coefficients respectively. The mass formula coefficients are represented as

$$\bar{\alpha} = \alpha \exp(j\theta_\alpha) \quad \bar{\gamma} = \gamma \exp(j\theta_\gamma) \quad (237)$$

$$\bar{\delta} = \delta \exp(j\theta_\delta) \quad \bar{\beta} = \beta \exp(j\theta_\beta) \quad (238)$$

$$\bar{E}_{\text{pair}} = E_{\text{pair}} \exp(j\theta_{E_{\text{pair}}}) \quad \bar{E}_{\text{shell}} = E_{\text{shell}} \exp(j\theta_{E_{\text{shell}}}) \quad (239)$$

where the complex number pairing energy is written as

$$\bar{E}_{\text{pair}} = P\bar{\rho}/\bar{a}^{3/4} \quad (240)$$

where the complex number pairing energy coefficient is written as

$$\bar{\rho} = \rho \exp(j\theta_\rho) \quad (241)$$

so that

$$E_{\text{pair}} = P\rho/a^{3/4} \quad \theta_{E_{\text{pair}}} = \theta_\rho - 3/4\theta_a \quad (242)$$

The average complex number binding energy per nucleon  $\bar{\epsilon} = \bar{B}/\bar{a}$  is written as

$$\begin{aligned} \bar{\epsilon} &= \bar{\epsilon}_v - \bar{\epsilon}_s - \bar{\epsilon}_c - \bar{\epsilon}_{\text{sym}} + \bar{\epsilon}_{\text{pair}} + \bar{\epsilon}_{\text{shell}} \\ &= \bar{\alpha} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}\bar{z}^2/\bar{a}^{4/3} - \bar{\beta}[(\bar{n} - \bar{z})/\bar{a}]^2 + P\bar{\rho}/\bar{a}^{7/4} + \bar{\epsilon}_{\text{shell}} \end{aligned} \quad (243)$$

where  $\bar{\epsilon}_v$ ,  $\bar{\epsilon}_s$ ,  $\bar{\epsilon}_c$ ,  $\bar{\epsilon}_{\text{sym}}$ ,  $\bar{\epsilon}_{\text{pair}}$  and  $\bar{\epsilon}_{\text{shell}}$  = complex number average volume, surface, Coulomb, symmetry, pairing and shell energies per nucleon respectively.

The complex number neutron excess that appears in the symmetry energy terms in equations (236) and (243) is written for actinide nuclei exactly as

$$\bar{\xi} = \xi \exp(j\theta_\xi) = \bar{n} - \bar{z} = \bar{a} - 2\bar{z} \quad (244)$$

where  $\bar{n}$ ,  $\bar{z}$  and  $\bar{a}$  are given by equations (43) through (45) and are related by equation (62). Then for actinide nuclei

$$\begin{aligned} \xi^2 &= n^2 + z^2 - 2zn \cos(\theta_n - \theta_z) \\ &= N^2 \sec^2 \theta_n + Z^2 \sec^2 \theta_z - 2ZN \sec \theta_z \sec \theta_n \cos(\theta_n - \theta_z) \end{aligned} \quad (245)$$

$$\begin{aligned} \tan \theta_\xi &= (n \sin \theta_n - z \sin \theta_z) / (n \cos \theta_n - z \cos \theta_z) \\ &= (N \tan \theta_n - Z \tan \theta_z) / (N - Z) \end{aligned} \quad (246)$$

For the approximation  $\theta_z \sim \theta_n$ , which is valid near the valley of beta stability, it follows from equations (245) and (246) that

$$\xi \sim n - z = N \sec \theta_n - Z \sec \theta_z \sim (N - Z) \sec \theta_z \quad (247)$$

$$\theta_\xi \sim \theta_z \sim \theta_n \quad (248)$$

Combining equations (243) and (244) gives the exact equation

$$\bar{\epsilon} = \bar{\alpha} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}\bar{z}^2/\bar{a}^{4/3} - \bar{\beta}(\bar{\xi}/\bar{a})^2 + P\bar{\rho}/\bar{a}^{7/4} + \bar{\epsilon}_{\text{shell}} \quad (249)$$

It follows from equation (244) that the following exact equation is valid for actinide nuclei

$$\begin{aligned}
\bar{\xi}/\bar{a} &= \xi/a \exp[j(\theta_{\xi} - \theta_a)] \\
&= 1 - 2\bar{z}/\bar{a} \\
&= 1 - 2(Z \sec \theta_z)/(A \sec \theta_a) \exp[j(\theta_z - \theta_a)]
\end{aligned} \tag{250}$$

Combining equations (249) and (250) gives the following exact expression for the average energy per nucleon for actinide nuclei

$$\bar{E} = \bar{\alpha} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}\bar{z}^2/\bar{a}^{4/3} - \bar{\beta}(1 - 2\bar{z}/\bar{a})^2 + P\bar{\rho}/\bar{a}^{7/4} + \bar{E}_{\text{shell}} \tag{251}$$

which is a simple complex number generalization of the standard scalar result.

Equation (236) giving the complex number binding energy of an actinide nucleus can be written as

$$\begin{aligned}
\bar{B} &= \alpha A \sec \theta_a \exp[j(\theta_{\alpha} + \theta_a)] - \gamma A^{2/3} \sec^{2/3} \theta_a \exp[j(\theta_{\gamma} + 2/3\theta_a)] \\
&- \delta Z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \exp[j(\theta_{\delta} + 2\theta_z - 1/3\theta_a)] \\
&- \beta \xi^2 A^{-1} \sec^{-1} \theta_a \exp[j(\theta_{\beta} + 2\theta_{\xi} - \theta_a)] \\
&+ P\rho A^{-3/4} \sec^{-3/4} \theta_a \exp[j(\theta_{\rho} - 3/4\theta_a)] + E_{\text{shell}} \exp(j\theta_{\text{Eshell}})
\end{aligned} \tag{252}$$

As an approximation the phase angles of the terms in equation (252) are taken to be equal

$$\begin{aligned}
\theta_B &\sim \theta_{\alpha} + \theta_a \sim \theta_{\gamma} + 2/3\theta_a \\
&\sim \theta_{\delta} + 2\theta_z - 1/3\theta_a \sim \theta_{\beta} + 2\theta_{\xi} - \theta_a \\
&\sim \theta_{\rho} - 3/4\theta_a \sim \theta_{\text{Eshell}}
\end{aligned} \tag{253}$$

Then the magnitude of the binding energy for actinide nuclei is obtained from equation (252) to be approximately

$$\begin{aligned}
B &\sim \alpha A \sec \theta_a - \gamma A^{2/3} \sec^{2/3} \theta_a - \delta Z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \\
&- \beta \xi^2 A^{-1} \sec^{-1} \theta_a + P\rho A^{-3/4} \sec^{-3/4} \theta_a + E_{\text{shell}}
\end{aligned} \tag{254}$$

Equation (59) gives  $\theta_a = \theta_a(\theta_z, \theta_n, Z, A)$  so that in all further calculations it should be understood that  $\theta_a$  is not really an independent variable. The approximation  $\theta_z \sim \theta_n$  allows equation (254) to be written as

$$\begin{aligned}
B &\sim \alpha A \sec \theta_a - \gamma A^{2/3} \sec^{2/3} \theta_a - \delta Z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \\
&- \beta (N - Z)^2 A^{-1} \sec^2 \theta_z \sec^{-1} \theta_a + P\rho A^{-3/4} \sec^{-3/4} \theta_a + E_{\text{shell}}
\end{aligned} \tag{255}$$

The further approximation  $\theta_F \sim \theta_Z \sim \theta_n \sim \theta_a$  allows equation (253) and (254) to be written for actinide nuclei as

$$\begin{aligned}\theta_B &\sim \theta_\alpha + \theta_a \sim \theta_\gamma + 2/3\theta_a \\ &\sim \theta_\delta + 5/3\theta_a \sim \theta_\beta + \theta_a \\ &\sim \theta_\rho - 3/4\theta_a \sim \theta_{\text{Eshell}}\end{aligned}\quad (256)$$

$$\begin{aligned}B &\sim \alpha A \sec \theta_a - \gamma A^{2/3} \sec^{2/3} \theta_a - \delta Z^2 A^{-1/3} \sec^{5/3} \theta_a \\ &- \beta (N - Z)^2 A^{-1} \sec \theta_a + P \rho A^{-3/4} \sec^{-3/4} \theta_a + E_{\text{shell}}\end{aligned}\quad (257)$$

which are valid in the valley of beta stability. For actinides at the fission condition  $\theta_a \sim 2\theta_Z$  equation (64) gives the approximations  $\theta_n \sim 2.6\theta_Z$  and  $\theta_n \sim 1.3\theta_a$ .

Each term in the expression for the binding energy and average binding energy per nucleon of actinide nuclei will now be considered separately with the exception of the nuclear shell effects which are more complicated and are not considered in this paper.

#### a. Volume Energy Term for Actinide Nuclei.

The volume energy terms are written for actinide nuclei as

$$\begin{aligned}\bar{E}_V &= \bar{\alpha}\bar{a} = \alpha a \exp[j(\theta_\alpha + \theta_a)] \\ &= \alpha A \sec \theta_a \exp[j(\theta_\alpha + \theta_a)]\end{aligned}\quad (258)$$

$$\bar{\epsilon}_V = \bar{\alpha} = \alpha \exp(j\theta_\alpha) \quad (259)$$

The volume energy per nucleon  $\bar{\alpha}$  describes the energy per nucleon of infinite nuclear matter but evaluated at the central density of a nucleus.<sup>36</sup>

#### b. Surface Energy Term for Actinide Nuclei.

The surface energy terms are written as

$$\begin{aligned}\bar{E}_S &= \bar{\gamma}\bar{a}^{2/3} = \gamma a^{2/3} \exp[j(\theta_\gamma + 2/3\theta_a)] \\ &= \gamma A^{2/3} \sec^{2/3} \theta_a \exp[j(\theta_\gamma + 2/3\theta_a)]\end{aligned}\quad (260)$$

$$\begin{aligned}\bar{\epsilon}_S &= \bar{\gamma}/\bar{a}^{1/3} = \gamma a^{-1/3} \exp[j(\theta_\gamma - 1/3\theta_a)] \\ &= \gamma A^{-1/3} \sec^{-1/3} \theta_a \exp[j(\theta_\gamma - 1/3\theta_a)]\end{aligned}\quad (261)$$

The complex number surface energy coefficient can be written in terms of the central density of an atomic nucleus.<sup>36</sup>

c. Coulomb Energy Term for Actinide Nuclei.

The complex number Coulomb energy terms are written as

$$\begin{aligned}\bar{E}_c &= \bar{\delta} z^2 / \bar{a}^{-1/3} = \delta z^2 a^{-1/3} \exp[j(\theta_\delta + 2\theta_z - 1/3\theta_a)] \\ &= \delta z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \exp[j(\theta_\delta + 2\theta_z - 1/3\theta_a)]\end{aligned}\quad (262)$$

$$\begin{aligned}\bar{E}_c &= \bar{\delta} z^2 / \bar{a}^{-4/3} = \delta z^2 a^{-4/3} \exp[j(\theta_\delta + 2\theta_z - 4/3\theta_a)] \\ &= \delta z^2 A^{-4/3} \sec^2 \theta_z \sec^{-4/3} \theta_a \exp[j(\theta_\delta + 2\theta_z - 4/3\theta_a)]\end{aligned}\quad (263)$$

The complex number Coulomb energy coefficient can be written as a generalization of the standard scalar result<sup>28-36</sup>

$$\begin{aligned}\bar{\delta} &= 3/5(e^2/\bar{b}) = 0.863/\bar{b} \quad \text{MeV} \\ &= (0.863/1.523)\bar{k}_c \quad \text{MeV}\end{aligned}\quad (264)$$

where  $\bar{b}$  = complex number radius parameter defined in equation (226). Equation (264) is equivalent to

$$\delta = 0.863/b = (0.863/1.523)k_c \quad \text{MeV} \quad (265)$$

$$\theta_\delta = -\theta_b = \theta_{kc} \quad (266)$$

For simplicity it will be assumed that the wave number  $k_c$  of the central density of an atomic nucleus is approximately equal to the wave number of infinite nuclear matter  $k_F$ , so that

$$k_c \sim k_F = 1.35 \text{ fm}^{-1} \quad (267)$$

but in fact  $k_c$  is slightly larger or smaller than  $k_F$  due to Coulomb and surface forces.<sup>36</sup>

d. Symmetry Energy Term for Actinide Nuclei.

The complex number symmetry energy terms are written as

$$\begin{aligned}\bar{E}_{\text{sym}} &= \bar{\beta} \xi^2 / \bar{a} = \beta \xi^2 a^{-1} \exp[j(\theta_\beta + 2\theta_\xi - \theta_a)] \\ &= \beta \xi^2 A^{-1} \sec^{-1} \theta_a \exp[j(\theta_\beta + 2\theta_\xi - \theta_a)]\end{aligned}\quad (268)$$

$$\begin{aligned}\bar{\epsilon}_{\text{sym}} &= \bar{B}(\bar{\xi}/\bar{A})^2 = \beta \xi^2 A^{-2} \exp[j(\theta_\beta + 2\theta_\xi - 2\theta_a)] \\ &= \beta \xi^2 A^{-2} \sec^{-2}\theta_a \exp[j(\theta_\beta + 2\theta_\xi - 2\theta_a)]\end{aligned}\quad (269)$$

where  $\bar{\xi}$ ,  $\xi$  and  $\theta_\xi$  are given by equations (244) through (246) respectively. Equations (268) and (269) can be simplified by assuming the approximation  $\theta_z \sim \theta_n$ , then equations (245) and (246) give

$$\xi \sim (N - Z) \sec \theta_z \quad \theta_\xi \sim \theta_z \sim \theta_n \quad (270)$$

and equations (268) and (269) become

$$\bar{E}_{\text{sym}} \sim \beta (N - Z)^2 A^{-1} \sec^{-1}\theta_a \sec^2\theta_z \exp[j(\theta_\beta + 2\theta_z - \theta_a)] \quad (271)$$

$$\bar{\epsilon}_{\text{sym}} \sim \beta (N - Z)^2 A^{-2} \sec^{-2}\theta_a \sec^2\theta_z \exp[j(\theta_\beta + 2\theta_z - 2\theta_a)] \quad (272)$$

As a further approximation let  $\theta_z \sim \theta_n \sim \theta_a$ , which follows from equation (59) for  $\theta_z \sim \theta_n$ , then equations (271) and (272) become

$$\bar{E}_{\text{sym}} \sim \beta (N - Z)^2 A^{-1} \sec \theta_a \exp[j(\theta_\beta + \theta_a)] \quad (273)$$

$$\bar{\epsilon}_{\text{sym}} \sim \beta (N - Z)^2 A^{-2} \exp(j\theta_\beta) \quad (274)$$

Equations (273) and (274) are valid in the vicinity of the valley of beta stability where  $\theta_z \sim \theta_n \sim \theta_a$ . The complex number symmetry energy coefficient can be written in terms of the central density of an atomic nucleus.<sup>36</sup>

#### e. Pairing Energy Term for Actinide Nuclei.

The complex number pairing terms are

$$\begin{aligned}\bar{E}_{\text{pair}} &= P\bar{\rho}/\bar{a}^{3/4} = P\rho a^{-3/4} \exp[j(\theta_\rho - 3/4\theta_a)] \\ &= P\rho A^{-3/4} \sec^{-3/4}\theta_a \exp[j(\theta_\rho - 3/4\theta_a)]\end{aligned}\quad (275)$$

$$\begin{aligned}\bar{\epsilon}_{\text{pair}} &= P\bar{\rho}/\bar{a}^{7/4} = P\rho a^{-7/4} \exp[j(\theta_\rho - 7/4\theta_a)] \\ &= P\rho A^{-7/4} \sec^{-7/4}\theta_a \exp[j(\theta_\rho - 7/4\theta_a)]\end{aligned}\quad (276)$$

The shell energy term is more complicated.

#### c. Measured Binding Energies of Actinide Nuclei Located in an External Field.

The real and imaginary parts of equation (236) are given by

$$\begin{aligned}
 B \cos \theta_B = & \alpha A \sec \theta_a \cos(\theta_\alpha + \theta_a) - \gamma A^{2/3} \sec^{2/3} \theta_a \cos(\theta_\gamma + 2/3 \theta_a) \quad (277) \\
 & - \delta Z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \cos(\theta_\delta + 2\theta_z - 1/3 \theta_a) \\
 & - \beta \xi^2 A^{-1} \sec^{-1} \theta_a \cos(\theta_\beta + 2\theta_\xi - \theta_a) \\
 & + P\rho A^{-3/4} \sec^{-3/4} \theta_a \cos(\theta_\rho - 3/4 \theta_a) + E_{\text{shell}} \cos \theta_{\text{Eshell}}
 \end{aligned}$$

$$\begin{aligned}
 B \sin \theta_B = & \alpha A \sec \theta_a \sin(\theta_\alpha + \theta_a) - \gamma A^{2/3} \sec^{2/3} \theta_a \sin(\theta_\gamma + 2/3 \theta_a) \quad (278) \\
 & - \delta Z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \sin(\theta_\delta + 2\theta_z - 1/3 \theta_a) \\
 & - \beta \xi^2 A^{-1} \sec^{-1} \theta_a \sin(\theta_\beta + 2\theta_\xi - \theta_a) \\
 & + P\rho A^{-3/4} \sec^{-3/4} \theta_a \sin(\theta_\rho - 3/4 \theta_a) + E_{\text{shell}} \sin \theta_{\text{Eshell}}
 \end{aligned}$$

Equations (277) and (278) immediately determine  $B$  and  $\theta_B$ . The measured binding energy is just the real part of the complex number binding energy, so that

$$B_m = \alpha_m A - \gamma_m A^{2/3} - \delta_m Z^2 A^{-1/3} - \beta'_m \xi^2 A^{-1} + E_{\text{pair}}^m + E_{\text{shell}}^m \quad (279)$$

where for actinide nuclei

$$\alpha_m = \alpha \sec \theta_a \cos(\theta_\alpha + \theta_a) \quad (280)$$

$$\gamma_m = \gamma \sec^{2/3} \theta_a \cos(\theta_\gamma + 2/3 \theta_a) \quad (281)$$

$$\delta_m = \delta \sec^2 \theta_z \sec^{-1/3} \theta_a \cos(\theta_\delta + 2\theta_z - 1/3 \theta_a) \quad (282)$$

$$\beta'_m = \beta \sec^{-1} \theta_a \cos(\theta_\beta + 2\theta_\xi - \theta_a) \quad (283)$$

$$E_{\text{pair}}^m = P\rho A^{-3/4} \sec^{-3/4} \theta_a \cos(\theta_\rho - 3/4 \theta_a) \quad (284)$$

where  $\xi = \xi(Z, N, \theta_z, \theta_n)$  and is defined by equation (245), and  $\theta_\xi = \theta_\xi(Z, N, \theta_z, \theta_n)$  is defined by equation (246). The internal phase angle  $\theta_a$  is given by equation (59) to be  $\theta_a = \theta_a(Z, N, \theta_z, \theta_n)$ .

If the small angle approximation  $\theta_\xi \sim \theta_z \sim \theta_n$  is assumed, then the approximations in equation (270) and (271) allow equation (279) to be written as

$$B_m = \alpha_m A - \gamma_m A^{2/3} - \delta_m Z^2 A^{-1/3} - \beta_m (N - Z)^2 A^{-1} + E_{\text{pair}}^m + E_{\text{shell}}^m \quad (285)$$



where now

$$\beta_m = \beta \sec^2 \theta_z \sec^{-1} \theta_a \cos(\theta_\beta + 2\theta_z - \theta_a) \quad (286)$$

If the further small angle approximation  $\theta_z \sim \theta_a$  is made, then equation (285) is the measured binding energy with

$$\alpha_m = \alpha \sec \theta_a \cos(\theta_\alpha + \theta_a) \quad (287)$$

$$\gamma_m = \gamma \sec^{2/3} \theta_a \cos(\theta_\gamma + 2/3\theta_a) \quad (288)$$

$$\delta_m = \delta \sec^{5/3} \theta_a \cos(\theta_\delta + 5/3\theta_a) \quad (289)$$

$$\beta_m = \beta \sec \theta_a \cos(\theta_\beta + \theta_a) \quad (290)$$

which are useful for nuclei near the valley of beta stability. The measured values of the symmetry energy coefficients are<sup>28,29</sup>

$$\alpha_m = 15.5 \quad \gamma_m = 17.2 \quad \delta_m = 0.698 \quad \beta_m = 23.3 \text{ MeV} \quad (291)$$

Equations (287) through (290) show that

$$\alpha_m < \alpha \quad \gamma_m < \gamma \quad \delta_m < \delta \quad \beta_m < \beta \quad (292)$$

The values of the nuclear mass formula parameters  $\theta_z$ ,  $\theta_n$ ,  $\theta_a$ ,  $\alpha$ ,  $\theta_\alpha$ ,  $\gamma$ ,  $\theta_\gamma$ ,  $\delta$ ,  $\theta_\delta$ ,  $\beta$  and  $\theta_\beta$  can be obtained by fitting the real part of the complex number binding energy given by equation (279) to the measured values of the atomic masses of the actinide elements. A simplified procedure uses the approximation  $\theta_z \sim \theta_n$  and equation (285) for the fit to atomic masses. Expressions for the atomic masses of the elements will now be considered.

#### D. Masses of Actinide Atoms Located in an Electromagnetic or Gravitational Field.

The conventional relationship between atomic mass and nuclear binding energy is written as<sup>29</sup>

$$M = Zm_H + Nm_n - B \quad (293)$$

where  $M$  = atomic mass of an element,  $m_H$  = mass of hydrogen atom and  $m_n$  = neutron mass. In the presence of an external field the atomic mass is a complex number in an internal space and is given by<sup>26</sup>

$$\bar{M} = \bar{Z}m_H + \bar{N}m_n - \bar{B} \quad (294)$$

where the complex number atomic mass is written as

$$\bar{M} = M \exp(j\theta_M) \quad (295)$$

Using equations (51) and (52) allows the real and imaginary parts of equation (294) to be written as

$$M \cos \theta_M = G \quad (296)$$

$$M \sin \theta_M = F \quad (297)$$

where for actinide nuclei

$$G = m_H Z + m_n N - B \cos \theta_B \quad (298)$$

$$F = m_H Z \tan \theta_z + m_n N \tan \theta_n - B \sin \theta_B \quad (299)$$

Equations (296) through (299) can be used to obtain  $M$  and  $\theta_M$  as

$$\tan \theta_M = F/G \quad (300)$$

$$M^2 = F^2 + G^2 \quad (301)$$

The measured atomic mass for the actinide elements is given by equations (296) and (298) which can be rewritten as

$$M_m = m_H Z + m_n N - B_m \quad (302)$$

where  $B_m$  is given by equations (279) or (285). The small angle approximation  $\theta_z \sim \theta_n \sim \theta_a \sim \theta_B$  combined with equations (294) through (302) gives for actinide nuclei

$$\theta_M \sim \theta_z \sim \theta_n \sim \theta_a \sim \theta_B \quad (303)$$

$$M \sim (m_H Z + m_n N) \sec \theta_z - B \quad (304)$$

where  $\theta_a$  is given by equation (63) within this approximation. Note that  $M$ ,  $M_m$  and  $\theta_M$  vary with the strength of the applied electromagnetic field because  $\theta_z = \theta_z(H)$  in equations (300) through (302), but the following intrinsic mass is a constant independent of the applied electromagnetic (or gravitational) field

$$m_H Z + m_n N = \text{constant} \quad (305)$$

and represents the universal law of the conservation of rest mass and baryon number.

The variation of the measured atomic mass and the magnitude of the atomic mass with the strength of the external electromagnetic field can be obtained from equations (301) and (302) by the following formulas

$$dM_m/dH = \partial M_m/\partial \theta_z d\theta_z/dH + \partial M_m/\partial \theta_n d\theta_n/dH + \partial M_m/\partial \theta_a d\theta_a/dH \quad (306)$$

$$dM/dH = \partial M/\partial \theta_z d\theta_z/dH + \partial M/\partial \theta_n d\theta_n/dH + \partial M/\partial \theta_a d\theta_a/dH \quad (307)$$

The internal phase angle  $\theta_a = \theta_a(\theta_z, \theta_n, Z, N)$  is given by equation (59) so that the derivative  $d\theta_a/dH$  can be evaluated as

$$d\theta_a/dH = \partial \theta_a/\partial \theta_z d\theta_z/dH + \partial \theta_a/\partial \theta_n d\theta_n/dH \quad (308)$$

Therefore equations (306) and (307) can be written as

$$\begin{aligned} dM_m/dH &= (\partial M_m/\partial \theta_z + \partial M_m/\partial \theta_a \partial \theta_a/\partial \theta_z) d\theta_z/dH \\ &+ (\partial M_m/\partial \theta_n + \partial M_m/\partial \theta_a \partial \theta_a/\partial \theta_n) d\theta_n/dH \end{aligned} \quad (309)$$

$$\begin{aligned} dM/dH &= (\partial M/\partial \theta_z + \partial M/\partial \theta_a \partial \theta_a/\partial \theta_z) d\theta_z/dH \\ &+ (\partial M/\partial \theta_n + \partial M/\partial \theta_a \partial \theta_a/\partial \theta_n) d\theta_n/dH \end{aligned} \quad (310)$$

where for example equation (302) gives for actinide nuclei

$$\partial M_m/\partial \theta_z = - \partial B_m/\partial \theta_z \quad (311)$$

$$\partial M_m/\partial \theta_n = - \partial B_m/\partial \theta_n \quad (312)$$

$$\partial M_m/\partial \theta_a = - \partial B_m/\partial \theta_a \quad (313)$$

where  $B_m$  is given by equation (279).

For the approximate case  $\theta_z \sim \theta_n \sim \theta_a$ , which follows from equation (59) for small arguments, it follows that

$$dM_m/dH = dM_m/d\theta_a d\theta_a/dH \quad (314)$$

$$dM/dH = dM/d\theta_a d\theta_a/dH \quad (315)$$

where equations (302) and (304) give for  $\theta_z \sim \theta_n \sim \theta_a$  and for actinide nuclei

$$dM_m/d\theta_a = - dB_m/d\theta_a \quad (316)$$

$$dM/d\theta_a = (m_H Z + m_N N) \sec \theta_a \tan \theta_a - dB/d\theta_a \quad (317)$$

where the approximate value of  $B_m$  given by equations (285) and (287) through (290) are used in conjunction with equation (316), while the approximate value of  $B$  given by equation (257) is used in conjunction with equation (317). Therefore from equation (285) and within the approximation  $\theta_z \sim \theta_n \sim \theta_a$ , the derivative in equation (316) is given by

$$dB_m/d\theta_a = d\alpha_m/d\theta_a A - d\gamma_m/d\theta_a A^{2/3} - d\delta_m/d\theta_a Z^2 A^{-1/3} \quad (318)$$

$$- d\beta_m/d\theta_a (N - Z)^2 A^{-1} + dE_{\text{pair}}^m/d\theta_a + dE_{\text{shell}}^m/d\theta_a$$

where from equations (287) through (290) for  $\theta_z \sim \theta_n \sim \theta_a$  it follows that for actinide nuclei

$$d\alpha_m/d\theta_a = -\alpha \sec \theta_a \cos(\theta_\alpha + \theta_a) [\tan(\theta_\alpha + \theta_a) - \tan \theta_a] \quad (319)$$

$$d\gamma_m/d\theta_a = -2/3\gamma \sec^{2/3} \theta_a \cos(\theta_\gamma + 2/3\theta_a) [\tan(\theta_\gamma + 2/3\theta_a) - \tan \theta_a] \quad (320)$$

$$d\delta_m/d\theta_a = -5/3\delta \sec^{5/3} \theta_a \cos(\theta_\delta + 5/3\theta_a) [\tan(\theta_\delta + 5/3\theta_a) - \tan \theta_a] \quad (321)$$

$$d\beta_m/d\theta_a = -\beta \sec \theta_a \cos(\theta_\beta + \theta_a) [\tan(\theta_\beta + \theta_a) - \tan \theta_a] \quad (322)$$

The derivative of the measured pairing energy that appears in equation (318) is obtained from equation (284) for actinide nuclei to be

$$dE_{\text{pair}}^m/d\theta_a = 3/4P\rho A^{-3/4} \sec^{-3/4} \theta_a \cos(\theta_\rho - 3/4\theta_a) [\tan(\theta_\rho - 3/4\theta_a) - \tan \theta_a] \quad (322A)$$

For the approximation  $\theta_z = \theta_n = \theta_a$ , the derivative of the magnitude of the binding energy for actinide nuclei that appears in equation (317) is obtained from equation (257) to be

$$dB/d\theta_a = \sec \theta_a \tan \theta_a [\alpha A - 2/3\gamma A^{2/3} \sec^{-1/3} \theta_a \quad (323)$$

$$- 5/3\delta Z^2 A^{-1/3} \sec^{2/3} \theta_a - \beta(N - Z)^2 A^{-1}$$

$$- 3/4P\rho A^{-3/4} \sec^{-1/4} \theta_a] + dE_{\text{shell}}/d\theta_a$$

From equations (316) through (323) it follows that for the actinides

$$dB_m/d\theta_a < 0 \quad dB/d\theta_a > 0 \quad (324)$$

$$dM_m/d\theta_a > 0 \quad dM/d\theta_a > 0 \quad (325)$$

For example, within the approximation  $\theta_z \sim \theta_n \sim \theta_a$  equations (317) and (323) give for actinide nuclei

$$dM/d\theta_a = \sec \theta_a \tan \theta_a [m_H Z + m_n N - \alpha A + 2/3\gamma A^{2/3} \sec^{-1/3} \theta_a \quad (326)$$

$$+ 5/3\delta Z^2 A^{-1/3} \sec^{2/3} \theta_a + \beta(N - Z)^2 A^{-1} + 3/4P\rho A^{-3/4} \sec^{-1/4} \theta_a]$$

$$- dE_{\text{shell}}/d\theta_a$$

which is a positive number because of the dominant contribution of the rest mass terms.

#### E. Valley of Beta Stability for Nuclei with Broken Internal Symmetries.

Radioactive beta decays require that the complex atomic number  $\bar{z}$  adjusts itself so as to minimize the binding energy of a nucleus given by equation (236) but subject to the constraints represented in equations (55) and (62).<sup>34</sup> Combining equations (62) and (236) gives after neglecting shell and pairing energy effects

$$\bar{B} = \bar{\alpha}\bar{a} - \bar{\gamma}\bar{a}^{2/3} - \bar{\delta}\bar{z}^2/\bar{a}^{1/3} - \bar{\beta}(\bar{a} - 2\bar{z})^2/\bar{a} \quad (327)$$

The minimum binding energy condition is

$$\partial\bar{B}/\partial\bar{z}|_{\bar{a}} = -2\bar{\delta}\bar{z}/\bar{a}^{1/3} + 4\bar{\beta}(\bar{a} - 2\bar{z})/\bar{a} = 0 \quad (328)$$

which gives using equation (62)

$$\bar{z}^{vs} = \bar{a}/2(1 + \bar{c}\bar{a}^{2/3})^{-1} \quad (329)$$

$$\bar{n}^{vs} = \bar{a}/2(1 + 2\bar{c}\bar{a}^{2/3})(1 - \bar{c}\bar{a}^{2/3})^{-1} \quad (330)$$

where vs = valley of beta stability, and where

$$\bar{c} = \bar{\delta}/(4\bar{\beta}) \quad c = \delta/(4\beta) \quad \theta_c = \theta_\delta - \theta_\beta \quad (331)$$

For medium weight atomic nuclei the following approximation to equations (329) and (330) can be used

$$\bar{z}^{vs} = \bar{a}/2(1 - \bar{c}\bar{a}^{2/3}) \quad (332)$$

$$\bar{n}^{vs} = \bar{a}/2(1 + \bar{c}\bar{a}^{2/3}) \quad (333)$$

where  $\bar{c}$  is given by equation (331).

For heavy nuclei the exact equation (329) must be used to calculate  $\bar{z}^{vs}$ . Equation (329) can be written as

$$\bar{z}^{vs} = (G + jF)/D \quad (334)$$

$$\tan \theta_z^{vs} = F/G \quad z^{vs} = (G^2 + F^2)^{1/2}/D \quad (335)$$

where

$$G = a/2 \cos \theta_a [1 + ca^{2/3} \cos(\theta_c + 2/3\theta_a)] + c/2 a^{5/3} \sin \theta_a \sin(\theta_c + 2/3\theta_a) \quad (336)$$

$$F = a/2 \sin \theta_a [1 + ca^{2/3} \cos(\theta_c + 2/3\theta_a)] - c/2 a^{5/3} \cos \theta_a \sin(\theta_c + 2/3\theta_a) \quad (337)$$

$$D = [1 + ca^{2/3} \cos(\theta_c + 2/3\theta_a)]^2 + c^2 a^{4/3} \sin^2(\theta_c + 2/3\theta_a) \quad (338)$$

Combining equations (46) and (336) through (338) gives for actinide nuclei

$$G = A/2 [1 + cA^{2/3} \sec^{2/3} \theta_a \cos(\theta_c + 2/3\theta_a)] \quad (339)$$

$$+ c/2A^{5/3} \sec^{2/3} \theta_a \tan \theta_a \sin(\theta_c + 2/3\theta_a)$$

$$F = A/2 \tan \theta_a [1 + cA^{2/3} \sec^{2/3} \theta_a \cos(\theta_c + 2/3\theta_a)] \quad (340)$$

$$- c/2A^{5/3} \sec^{2/3} \theta_a \sin(\theta_c + 2/3\theta_a)$$

$$D = [1 + cA^{2/3} \sec^{2/3} \theta_a \cos(\theta_c + 2/3\theta_a)]^2 \quad (341)$$

$$+ c^2 A^{4/3} \sec^{4/3} \theta_a \sin^2(\theta_c + 2/3\theta_a)$$

If  $A$  and  $\theta_a$  are taken to be the known quantities, then equation (54) involves five unknown quantities  $W$ ,  $Z$ ,  $N$ ,  $\theta_z$  and  $\theta_n$ . The complex number equation (54) and the scalar equation (55) supply three equations for determining the five unknown quantities. The complex number valley of beta stability equation (328) supplies two additional equations and so a complete solution is possible. From equations (335) and (339) through (341) that describe the valley of beta stability for the general case of nuclei of arbitrary size it follows that in general for actinide nuclei  $W = 0$  and

$$\theta_z^{vs} = \theta_z^{vs}(A, \theta_a) \quad Z^{vs} = Z^{vs}(A, \theta_a) \quad (342)$$

Combining equations (46) and (355) gives for actinide nuclei

$$Z^{vs} = z^{vs} \cos \theta_z^{vs} = G/D \quad (343)$$

$$N^{vs} = A - Z^{vs} = A - G/D \quad (344)$$

The value of  $\theta_n^{vs}$  for the valley stability of actinide nuclei is obtained exactly from equation (62) which is written in the form with  $W = 0$

$$\bar{n}^{vs} = \bar{a} - \bar{z}^{vs} \quad (345)$$

from which it follows in accordance with equation (60) that for actinide nuclei

$$\tan \theta_n^{vs} = (A \tan \theta_a - Z^{vs} \tan \theta_z^{vs}) / (A - Z) \quad (346)$$

where  $\theta_z^{vs}$  and  $Z^{vs}$  are given by equations (335) and (343). The measured values of the atomic number, neutron number and atomic mass number in the valley of beta stability are given for the actinides by

$$z_m^{vs} = Z^{vs} \quad n_m^{vs} = N^{vs} \quad a_m = A \quad (347)$$

which are integers.

**5. SUPPRESSION OF LOW ENERGY BINARY FISSION OF THE FISSIONABLE ACTINIDES BY AN EXTERNAL FIELD.** This section determines the conditions necessary for the inhibition of spontaneous or thermal neutron induced binary fission in fissionable actinide nuclei that are located in an electromagnetic or gravitational field. The Bohr-Wheeler fission instability condition is generalized to the case of actinide nuclei located in an external field. This condition is used to determine the critical internal phase angle of the atomic number which corresponds to the suppression of spontaneous or thermal neutron induced binary fission by an external field. An expression for the corresponding critical static magnetic field that is required to suppress binary fission in the fissionable actinides is presented and numerical values are obtained for this critical static magnetic field in terms of the fissionability parameters for several actinide nuclei.

**A. Bohr-Wheeler Fission Condition for Actinide Nuclei in an Electromagnetic Field.**

The standard Bohr-Wheeler analysis for spontaneous or thermal neutron induced nuclear fission utilizes the fissionability parameter which is defined by<sup>39-44</sup>

$$\chi = (Z^2/A)(\kappa\gamma/\delta)^{-1} \quad (348)$$

and the spontaneous and thermal neutron induced fission condition is written as<sup>39-44</sup>

$$\chi \geq 1 \quad Z^2/A \geq \kappa\gamma/\delta$$

where  $\gamma$  and  $\delta$  = surface and Coulomb energy coefficients that appear in the liquid drop nuclear mass formula treated in Section 4, and where theoretically for spontaneous fission

$$\kappa = g/h = 2 \quad (350)$$

where  $g = 2/5$  and  $h = 1/5$  are the second order series expansion coefficients of the surface and Coulomb energies respectively when these terms are expanded in terms of an ellipsoidal deformation parameter.<sup>39-44</sup> The values of  $\kappa$ ,  $\gamma$  and  $\delta$  along with the other mass formula parameters are determined empirically.<sup>39-44</sup> The values of  $\kappa$  are different for spontaneous and for thermal neutron induced fission, and in fact  $\kappa$  is dependent on the energy of the incident neutrons.<sup>39-44</sup> For thermal neutron induced fission<sup>39-44</sup>

$$\kappa \sim 1.471 \quad (351)$$

Choosing  $\gamma = 17.2$  MeV and  $\delta = 0.698$  MeV yields the following fission conditions for a zero value of the externally applied field<sup>39-44</sup>

$$Z^2/A \geq 49.28 \quad \text{spontaneous fission} \quad (352)$$

$$Z^2/A \geq 36.25 \quad \text{thermal neutron induced fission} \quad (353)$$

These inequalities show that, loosely speaking, only the actinides and trans-actinides can undergo spontaneous or thermal neutron induced fission, but not all of these heavy elements undergo fission. Within this group of heavy elements the more neutron rich isotopes tend to be more stable against fission, for example  $^{238}\text{U}$  is stable against thermal neutron induced fission but  $^{235}\text{U}$  is fissile. The empirical value of  $\kappa$  that describes thermal neutron induced fission will depend on the values selected for the mass formula parameters  $\gamma$  and  $\delta$ . In general  $\kappa$  can be taken to be a decreasing function of the kinetic energy of the incident neutrons. The fission criteria presented above ignore all shell structure effects and are therefore approximate relations which show only general behavior and for which counterexamples can always be found in the border region between fissile and non-fissile nuclei.

The generalization of equation (349) to the case of atomic nuclei located in an electromagnetic or gravitational field, which breaks the symmetry of the atomic number, neutron number and atomic mass number, can be written as<sup>26</sup>

$$\bar{z}^2/\bar{a} > \kappa\bar{\gamma}/\bar{\delta} \quad (354)$$

where  $\bar{z}$ ,  $\bar{a}$ ,  $\bar{\gamma}$  and  $\bar{\delta}$  are given by equations (43), (45), (237) and (238) respectively. The fission instability boundary is given by

$$\bar{z}^2/\bar{a} = \kappa\bar{\gamma}/\bar{\delta} \quad (355)$$

or equivalently the two scalar fission stability boundary conditions are<sup>26</sup>

$$z^2/a = \kappa\gamma/\delta \quad (356)$$

$$\theta_a = 2\theta_z - \theta_\gamma + \theta_\delta \quad (357)$$

Therefore in an external field the internal phase angles of the atomic number, atomic mass number, surface energy coefficient and the Coulomb energy coefficient enter into the fission instability condition. Equations (356) and (357) will be solved to determine the critical value of  $\theta_z$  that is required to suppress spontaneous or thermal neutron induced binary fission in the actinides by the application of an external electromagnetic field.

#### B. Critical Value of the Internal Phase Angle of the Atomic Number that is Required to Suppress Low Energy Binary Fission of Fissile Actinide Nuclei Located in an Electromagnetic Field.

Combining equation (46) with equations (356) and (357) gives the binary fission instability boundary for actinide nuclei located in an external field as

$$\begin{aligned} z^2/A &= \kappa\gamma/\delta \sec \theta_a \sec^{-2} \theta_z \\ &= \kappa\gamma/\delta \sec(2\theta_z + \theta_\delta - \theta_\gamma) \sec^{-2} \theta_z \end{aligned} \quad (358)$$

Equation (358) must be solved for  $\theta_z$ . This can be done by noting that simple trigonometry gives<sup>26</sup>



$$\cos(2\theta_z + \theta_\delta - \theta_\gamma) \cos^{-2}\theta_z = (1 - \rho^2) \cos(\theta_\gamma - \theta_\delta) + 2\rho \sin(\theta_\gamma - \theta_\delta) \quad (359)$$

where

$$\rho = \tan \theta_z \quad (360)$$

Equation (358) can then be written as a quadratic equation

$$ap^2 + bp + c = 0 \quad (361)$$

where

$$a = \cos(\theta_\gamma - \theta_\delta) \quad (362)$$

$$b = 2 \sin(\theta_\delta - \theta_\gamma) = -2 \sin(\theta_\gamma - \theta_\delta) \quad (363)$$

$$c = (\kappa\gamma/\delta)(A/Z^2) - \cos(\theta_\gamma - \theta_\delta) = \chi^{-1} - \cos(\theta_\gamma - \theta_\delta) \quad (364)$$

where the fissility parameter  $\chi$  is given by equation (40). Then the critical angle of the atomic number for binary fission suppression  $\theta_z^S$  is given by

$$\begin{aligned} \tan \theta_z^S &= \tan(\theta_\gamma - \theta_\delta) \pm \sec(\theta_\gamma - \theta_\delta) [1 - (\kappa\gamma/\delta)(A/Z^2) \cos(\theta_\gamma - \theta_\delta)]^{1/2} \\ &= \tan(\theta_\gamma - \theta_\delta) \pm \sec(\theta_\gamma - \theta_\delta) [1 - \chi^{-1} \cos(\theta_\gamma - \theta_\delta)]^{1/2} \end{aligned} \quad (365)$$

and from equation (357) the corresponding critical angle of the atomic mass number for fission suppression  $\theta_a^S$  is given by

$$\theta_a^S = 2\theta_z^S + \theta_\delta - \theta_\gamma \quad (366)$$

The angles  $\theta_z^S$  and  $\theta_a^S$  are the critical values of the phase angles  $\theta_z$  and  $\theta_a$  that are required to suppress thermal neutron induced binary fission in an actinide nucleus. In other words,  $\theta_z \geq \theta_z^S$  and  $\theta_a \geq \theta_a^S$  are required for binary fission suppression in an actinide nucleus by the presence of an external field. Figure 1 gives  $\theta_z^S$  and Figure 2 gives  $\theta_a^S$  in terms of the fissility parameter for actinide nuclei with the choice of phase angle values  $\theta_\gamma = 0.4\pi$  and  $\theta_\delta = 0.1\pi$  whose values are selected only for demonstration purposes.

Equation (365) is the equation for the instability boundary for the suppression of thermal neutron induced fission of an actinide nucleus ( $Z, A$ ), and is valid for

$$\kappa\gamma/\delta \cos(\theta_\gamma - \theta_\delta) \leq Z^2/A \leq \infty \quad (367)$$

or in terms of the fissility parameter

$$\cos(\theta_\gamma - \theta_\delta) \leq \chi \leq \infty \quad (368)$$

For  $\chi = \infty$  equation (365) becomes

$$\tan \theta_{z^\infty}^{S\pm} = \tan(\theta_\gamma - \theta_\delta) \pm \sec(\theta_\gamma - \theta_\delta) \quad (369)$$

Simple trigonometry then gives

$$\theta_{z^\infty}^{S+} = \pi/4 + 1/2(\theta_\gamma - \theta_\delta) \quad (370)$$

$$\theta_{z^\infty}^{S-} = -\pi/4 + 1/2(\theta_\gamma - \theta_\delta) \quad (371)$$

The common value of  $\theta_z^S$  and  $\theta_a^S$  that occurs at  $\chi = \cos(\theta_\gamma - \theta_\delta)$ , which is the minimum value allowed for the fissility parameter in equation (365), is given by

$$\theta_z^{Sc} = \theta_a^{Sc} = \theta_\gamma - \theta_\delta \quad (372)$$

The ranges of variation of  $\theta_z^S$  and  $\theta_a^S$  over both the positive and negative modes are given by

$$\theta_{z^\infty}^{S-} \leq \theta_z^S \leq \theta_{z^\infty}^{S+} \quad (373)$$

$$-\pi/2 \leq \theta_a^S \leq \pi/2$$

as shown in Figures 1 and 2 for  $\chi \geq \cos(\theta_\gamma - \theta_\delta)$ . The ranges of variation of  $\theta_z^S$  and  $\theta_a^S$  corresponding to the positive angle mode for the actinides are given by

$$\theta_z^{Sc} \leq \theta_z^S \leq \theta_{z^\infty}^{S+} \quad (375)$$

$$\theta_a^{Sc} \leq \theta_a^S \leq \pi/2 \quad (376)$$

while the ranges of variation of  $\theta_z^S$  and  $\theta_a^S$  for the negative angle mode for the actinides are given by

$$\theta_{z^\infty}^{S-} \leq \theta_z^S \leq \theta_z^{Sc} \quad (377)$$

$$-\pi/2 \leq \theta_a^S \leq \theta_a^{Sc} \quad (378)$$

for  $\chi \geq \cos(\theta_\gamma - \theta_\delta)$  as shown in Figures 1 and 2.

The condition for the suppression of thermal neutron induced binary fission in the actinides by an external electromagnetic field can also be written in an alternative form to equation (365) as follows

$$\cos(\theta_\gamma - \theta_\delta) \leq \chi \leq F^{-1} \quad (379)$$

where  $\chi$  = fissility parameter, and where

$$F = (1 - \lambda^2) \sec(\theta_\gamma - \theta_\delta) \quad (380)$$

$$\lambda = [\tan \theta_z - \tan(\theta_\gamma - \theta_\delta)] \cos(\theta_\gamma - \theta_\delta) \quad (381)$$

where in general  $F \leq 1$ . Equations (365) and (379) are equivalent to the following binary fission suppression condition

$$\theta_z \geq \theta_z^S \quad (382)$$

where  $\theta_z^S$  is given by equation (365) in terms of  $\chi$ . Equation (379) gives the range of fissility parameters for which thermal neutron induced binary fission is suppressed by an external field which generates the phase angle  $\theta_z$ . It is clear from equations (379) and (382) that in the presence of an electromagnetic field thermal neutron induced binary fission can occur only for actinide nuclei for which

$$\chi \geq F^{-1} \quad \theta_z \leq \theta_z^S \quad (383)$$

Equation (383) is the Bohr-Wheeler binary fission condition for actinide nuclei in an external field for which  $F \leq 1$ . The phase angle  $\theta_z^S$  represents a boundary between the regions of thermal neutron induced binary fission and binary fission suppression of the actinides in an electromagnetic field. If the external field is shut off all of the internal phase angles have zero values and  $\lambda = 0$  and  $F = 1$  so that equation (379) reduces to the statement that under zero field conditions there are no nuclei for which binary fission is suppressed because the binary fission suppression range given by equation (379) shrinks to zero length about  $\chi = 1$ , and equation (383) reduces to the standard Bohr-Wheeler fission condition given by equation (349).

As a first approximation the condition  $\theta_\gamma = \theta_\delta$  can be taken in equation (366) and the phase angle condition for the suppression of spontaneous or thermal neutron induced binary fission in an external field is

$$\theta_a^S = 2\theta_z^S \sim 0.76\theta_n^S \quad \theta_n^S \sim 2.63\theta_z^S \quad \theta_\xi^S \sim 5.41\theta_z^S \quad (384)$$

Combining equations (358) and (384) gives the approximate binary fission suppression boundary for actinide nuclei in an external field as

$$Z^2/A = (\kappa_\gamma/\delta)(1 - \tan^2 \theta_z^S)^{-1} \quad (385)$$

$$\chi = (1 - \tan^2 \theta_z^S)^{-1} \quad (386)$$

or equivalently as

$$Z^2/A = (\kappa_\gamma/\delta)[1 - \tan^2(\theta_a^S/2)]^{-1} \quad (387)$$

$$\chi = [1 - \tan^2(\theta_a^S/2)]^{-1} \quad (388)$$

For the case of the approximation  $\theta_\gamma = \theta_\delta$  the condition for binary fission suppression in fissile actinide nuclei by an external field which induces a phase angle  $\theta_z$  in the atomic number is given by equations (379) and (382) with  $\lambda = \tan \theta_z$  as

$$1 \leq (Z^2/A)(\kappa\gamma/\delta)^{-1} \leq (1 - \tan^2 \theta_z)^{-1} \quad \theta_z \geq \theta_z^S \quad (389)$$

or equivalently

$$1 \leq \chi \leq (1 - \tan^2 \theta_z)^{-1} \quad \theta_z \geq \theta_z^S \quad (390)$$

which follows directly from equations (379) through (381) when  $\theta_\gamma = \theta_\delta$ . Equation (390) determines the nuclei region of suppressed binary fission within the approximation  $\theta_\gamma = \theta_\delta$ , and corresponds to the exact region of suppressed binary fission that is specified by equations (379) through (382). Within the approximation  $\theta_\gamma = \theta_\delta$  the condition for thermal neutron induced binary fission of the actinides in an electromagnetic field is obtained from equation (390) as

$$\chi \geq (1 - \tan^2 \theta_z)^{-1} \quad \theta_z \leq \theta_z^S \quad (391)$$

which follows from equation (383) for the case  $\theta_\gamma = \theta_\delta$ , and which reduces to the standard Bohr-Wheeler fission condition given by equation (349) for the case of zero external field. Equations (386) and (388) determine  $\theta_z^S$  and  $\theta_a^S$  in terms of  $\chi$  for the approximation  $\theta_\gamma = \theta_\delta$ .

Equations (385) and (386) show that within the approximation  $\theta_\gamma = \theta_\delta$  the internal phase angle of the atomic number that is required to suppress thermal neutron induced binary fission in an actinide nucleus  $(Z,A)$  is given by

$$\begin{aligned} \tan \theta_z^S &= \pm [1 - (\kappa\gamma/\delta)(A/Z^2)]^{1/2} \\ &= \pm (1 - \chi^{-1})^{1/2} \\ &= \pm [(\chi - 1)/\chi]^{1/2} \end{aligned} \quad (392)$$

where  $\chi \geq 1$ . As shown in Figure 3 equation (392) has positive and negative modes. Equation (392) can be obtained directly from the exact equation (365) by making the approximation  $\theta_\gamma = \theta_\delta$ . The approximation  $\theta_\gamma = \theta_\delta$  is made only for convenience because this eliminates the values of  $\theta_\gamma$  and  $\theta_\delta$  from the calculation of  $\theta_z^S$ . If this approximation is not made the values of  $\theta_\gamma$  and  $\theta_\delta$  must be obtained from fitting a liquid drop type of nuclear mass formula to measured atomic masses. Equation (392) can also be written as

$$\begin{aligned} \cos \theta_z^S &= [2 - (\kappa\gamma/\delta)(A/Z^2)]^{-1/2} \\ &= (2 - \chi^{-1})^{-1/2} \\ &= [(2\chi - 1)/\chi]^{-1/2} \end{aligned} \quad (393)$$

$$\begin{aligned}\sin \theta_z^S &= \pm \{[1 - (\kappa_Y/\delta)(A/Z^2)]/[2 - (\kappa_Y/\delta)(A/Z^2)]\}^{1/2} \\ &= \pm [(1 - \chi^{-1})/(2 - \chi^{-1})]^{1/2} \\ &= \pm [(\chi - 1)/(2\chi - 1)]^{1/2}\end{aligned}\quad (394)$$

Equations (384) and (392) can be written equivalently as

$$\begin{aligned}\theta_z^S &= \pm \tan^{-1}[1 - (\kappa_Y/\delta)(A/Z^2)]^{1/2} & \theta_n^S &\sim (2 + Z/N)\theta_z^S \\ &= \pm \tan^{-1}[(\chi - 1)/\chi]^{1/2}\end{aligned}\quad (395)$$

$$\begin{aligned}\theta_a^S &= \pm 2 \tan^{-1}[1 - (\kappa_Y/\delta)(A/Z^2)]^{1/2} & \theta_\xi^S &\sim 2\theta_z^S/(1 - Z/N) \\ &= \pm 2 \tan^{-1}[(\chi - 1)/\chi]^{1/2} & &= \theta_a^S/(1 - Z/N)\end{aligned}\quad (396)$$

and are plotted in Figures 3 and 4 for  $\theta_Y = \theta_\delta$  for actinide nuclei with  $\chi \geq 1$ . The approximate equations (392) through (396) are valid for  $\theta_Y = \theta_\delta$  and the following range of the fissility parameter

$$1 \leq \chi \leq \infty \quad (397)$$

Within the approximation  $\theta_Y = \theta_\delta$  the range of values of  $\theta_z^S$  for the positive and negative modes is

$$-\pi/4 \leq \theta_z^S \leq \pi/4 \quad (398)$$

as shown in Figure 3 for  $\chi \geq 1$ . The range of values of  $\theta_a^S$  for the positive and negative modes is

$$-\pi/2 \leq \theta_a^S \leq \pi/2 \quad (399)$$

as shown in Figure 4 for  $\chi \geq 1$ . Equations (398) and (399) correspond to the exact relations given in equations (373) and (374) respectively. For the positive angle modes the ranges of  $\theta_z^S$  and  $\theta_a^S$  are within the approximation  $\theta_Y = \theta_\delta$  given by

$$0 \leq \theta_z^S \leq \pi/4 \quad 0 \leq \theta_a^S \leq \pi/2 \quad (400)$$

while for the negative angle modes

$$-\pi/4 \leq \theta_z^S \leq 0 \quad -\pi/2 \leq \theta_a^S \leq 0 \quad (401)$$

which correspond to the general cases in equations (375) through (378). Values of the angle  $\theta_z^S$  given by equation (395) for various actinide nuclei appear in Table 1.

#### C. Determination of the Static Magnetic Field Required to Suppress Thermal Neutron Induced Binary Fission in Actinide Nuclei.

The value of the magnetic field required to suppress binary fission in the actinides is calculated by assuming a relationship between the magnetic field and the internal phase angle of the atomic number that is required to suppress binary fission that is analogous to equation (42), so that

$$H^S = K_{\theta z}^H \tan \theta_z^S \quad H = K_{\theta z}^H \tan \theta_z \quad (402)$$

where  $K_{\theta z}^H$  = static nuclear magnetic stiffness coefficient, and where  $\theta_z^S$  is given for the general case by equation (365). Therefore in general

$$H^S = H^S(K_{\theta z}^H, \theta_\gamma - \theta_\delta, \chi) \quad (403)$$

where the fissility parameter  $\chi$  is given by equation (348). If  $\theta_z^S$  is given by the approximate equation (392) the critical value of the magnetic field required to suppress binary fission in the actinides is given by

$$H^S = K_{\theta z}^H [(\chi - 1)/\chi]^{1/2} \quad (404)$$

with  $\chi \geq 1$ , so that within this approximation

$$H^S = H^S(K_{\theta z}^H, \chi) \quad (405)$$

and  $H^S$  depends on only two parameters. The corresponding value of the static magnetic induction required to suppress binary fission is written as

$$B^S = K_{\theta z}^B \tan \theta_z^S \quad B = K_{\theta z}^B \tan \theta_z \quad (406)$$

$$= K_{\theta z}^B [(\chi - 1)/\chi]^{1/2} \quad (407)$$

where

$$K_{\theta z}^B = \mu K_{\theta z}^H \quad (408)$$

where  $\mu$  = magnetic permeability of nuclear matter. The static nuclear magnetic stiffness coefficients have the values<sup>26</sup>

$$K_{\theta z}^B = 2.36 \times 10^{13} \text{ T} \quad (409)$$

$$K_{\theta z}^H = 1.88 \times 10^{19} \text{ coul/(m sec)} \quad (410)$$

where T = tesla. The value of the magnetic permeability is taken to be the vacuum value<sup>26</sup>

$$\mu = \mu_0 = 4\pi \times 10^{-7} \text{ kg m/coul}^2 \quad (411)$$

A complete discussion of the determination of the coefficients  $K_{\theta z}^H$  and  $K_{\theta z}^B$  has appeared in the literature.<sup>26</sup> As shown in equation (390) thermal neutron in-

duced binary fission in the actinides is suppressed by an electromagnetic field when  $\theta_z \geq \theta_z^S$  or equivalently when  $B \geq B^S$ .

It has been shown that fissile actinide nuclei in the presence of an electromagnetic field may under some conditions be cooled so as to inhibit binary fission by thermal neutrons. For actinide nuclei in an electromagnetic field the internal phase angle  $\theta_z$  of the atomic number is an increasing function of the magnetic induction  $B$  which can be represented by equation (406). When the strength of the static magnetic field exceeds a critical value required for binary fission suppression, which is given by equation (406) for the general case or by equation (407) within a simple approximation, the phase angle  $\theta_z$  exceeds a critical value  $\theta_z^S$  which is required for binary fission suppression so that

$$\theta_z \geq \theta_z^S \quad (412)$$

where  $\theta_z^S$  is given exactly by equation (365) and approximately by equation (395). The corresponding static magnetic induction field condition for binary fission suppression in the actinides is given by

$$B \geq B^S \quad (413)$$

where from equation (407) the following approximate result can be used

$$B^S = K_{\theta_z}^B [(\chi - 1)/\chi]^{1/2} \quad 1 \leq \chi \leq \infty \quad (414)$$

where  $\chi$  = fissility parameter of an actinide nucleus. Values of the static magnetic induction field  $B^S$  required to suppress thermal neutron induced binary fission appear in Table 1 for various actinide nuclei. The values of  $B^S$  are in the teratesla range which shows that the static magnetic field required for the inhibition of fission in the actinides is much too large for practical purposes. However, more reasonable results can be obtained with a properly tuned electromagnetic field in the form of  $\gamma$  rays.

**6. QUATERNARY FISSION OF  $\gamma$  RAY COOLED ACTINIDE NUCLEI.** This section suggests that  $\gamma$  ray induced quaternary fission can occur in fissile actinide nuclei in which the binary fission mode has been suppressed by the cooling effects of a  $\gamma$  ray field. Numerical values of the static magnetic induction field  $B^S$  required to suppress thermal neutron induced binary fission in the actinides were presented in Section 5, and were found to be too large for practical applications. This section gives the corresponding values of the dynamic magnetic induction field  $B_\gamma^S$  associated with the  $\gamma$  rays that are required to suppress binary fission in the actinide elements. Thermal neutron induced binary fission will be suppressed in the actinides such as  $^{235}\text{U}$  and  $^{239}\text{Pu}$  when these nuclei are cooled by a properly tuned bath of ambient  $\gamma$  rays. However each of the two sub-actinide lobes of the distorted actinide nucleus, which is in the  $\gamma$  ray cooled binary fission suppressed state, can undergo  $\gamma$  ray catalyzed thermal neutron induced binary fission with the result that the actinide nucleus experiences quaternary fission. Examples of thermal neutron induced quaternary fission reactions in  $\gamma$  ray cooled actinides are given. The photonuclear sum rule for these reactions is evaluated. The resulting clean fission process for actinide

nuclei such as  $^{235}\text{U}$  and  $^{239}\text{Pu}$  can be used to develop environmentally safe nuclear power reactors because the enhanced quaternary fission of these heavy elements produces relatively light nuclei as fission products. Two design concepts of  $\gamma$  ray cooled actinide quaternary fission nuclear reactors are presented.

#### A. Conditions for the Thermal Neutron Induced Quaternary Fission of $\gamma$ Ray Cooled Actinides.

First the thermal neutron induced binary fission mode for the fissile actinides must be suppressed. The suppression of binary fission in the actinides can be accomplished by a  $\gamma$  ray field which is tuned to the giant dipole resonance frequency of the actinide element that is to be used in a clean fission nuclear reactor. The magnitude of the magnetic induction field of the  $\gamma$  rays is obtained by first noting that at the resonance frequency the dynamic nuclear magnetic stiffness coefficient is given by<sup>26</sup>

$$K_{\theta z}^{B\gamma} = \zeta_r K_{\theta z}^B \quad \text{tesla} \quad (415)$$

$$= \sigma A^{-1/6} K_{\theta z}^B \quad (416)$$

where  $K_{\theta z}^{B\gamma}$  = dynamic nuclear magnetic stiffness coefficient and where<sup>26</sup>

$$\sigma = 3.054 \times 10^{-9} \quad (417)$$

Values of  $\zeta_r$  and  $K_{\theta z}^{B\gamma}$  appear in Table 2 for various actinide nuclei. For binary fission suppression in the actinides using  $\gamma$  rays the dynamic magnetic induction field  $B_\gamma$  of the  $\gamma$  rays and the corresponding phase angle  $\theta_z$  of the atomic number must satisfy

$$B_\gamma \geq B_\gamma^S \quad \theta_z \geq \theta_z^S \quad (418)$$

where

$$B_\gamma = K_{\theta z}^{B\gamma} \tan \theta_z \quad B_\gamma^S = K_{\theta z}^{B\gamma} \tan \theta_z^S \quad (419)$$

where  $\theta_z^S$  is given by equation (365) or (395). For the choice of the approximate values of  $\theta_z^S$  given by equation (395) the critical dynamic magnetic induction field of the  $\gamma$  rays required for binary fission suppression in the actinides is given by

$$B_\gamma^S = K_{\theta z}^{B\gamma} [(\chi - 1)/\chi]^{1/2} \quad 1 \leq \chi \leq \infty \quad (420)$$

where  $\chi$  = fissility parameter for an actinide nucleus. Table 1 gives values of  $B_\gamma^S$  for selected actinide nuclei. The calculation of the giant dipole resonance frequency of atomic nuclei requires an estimation of the spring constant of an atomic nucleus.<sup>26</sup> The values of the nuclear spring constant  $k$ , resonance frequency  $f_r$  and resonance wavelength  $\lambda_r$  of the incident  $\gamma$  rays are given in Table 2 for selected actinide nuclei. For resonant incident  $\gamma$  rays the critical dynamic magnetic field strength  $H_\gamma^S$ , dynamic electric field strength  $E_\gamma^S$ , power



density  $P_Y^S$ , photon energy  $\epsilon_Y^S$ , photon flux density  $\Phi_Y^S$  and photon number density  $n_Y^S$  that are required to suppress binary fission in selected actinide nuclei are given in Table 3. The details of these calculations are given in Reference 26. The  $\gamma$  ray cooling of the binary fission process is a nuclear analog of the cooling of electrons and atoms by laser light.<sup>45-48</sup>

For a value of the  $\gamma$  ray magnetic induction field  $B_Y$  that satisfies the suppression condition for thermal neutron induced binary fission of actinide nuclei given by equation (418), an actinide nucleus may be considered to be composed of two lobes of subactinide nuclei a and b into which the actinide nucleus would fission were it not for the cooling effects of the  $\gamma$  ray electromagnetic field (see Figure 5). However, thermal neutron induced binary fission of the component subactinide nuclei lobes can occur if the electromagnetic field of the  $\gamma$  rays is sufficiently strong as to make the internal phase angles of the atomic numbers of the component subactinide nuclei a and b larger than their critical values required for  $\gamma$  ray catalyzed thermal neutron induced fission<sup>26</sup>

$$\theta_{za} \geq \theta_{za}^F \quad \theta_{zb} \geq \theta_{zb}^F \quad (421)$$

where  $\theta_{za}^F$  and  $\theta_{zb}^F$  are given exactly by equation (39) or approximately by equation (41) in terms of the fissility parameters  $\chi_a$  and  $\chi_b$  respectively of the two subactinide nuclei lobes. The net result is  $\gamma$  ray catalyzed thermal neutron induced quaternary fission of fissile actinide nuclei as shown in Figure 5. Accordingly, the magnetic induction field conditions for quaternary fission of  $\gamma$  ray cooled actinide nuclei using thermal neutrons are given by

$$\theta_{za} \geq \theta_{za}^F \quad \theta_{zb} \geq \theta_{zb}^F \quad \theta_z \geq \theta_z^S \quad (422)$$

$$B_Y \geq B_{aY}^F \quad B_Y \geq B_{bY}^F \quad B_Y \geq B_Y^S \quad (423)$$

where the critical magnetic field  $B_Y^S$  required for  $\gamma$  ray cooling of the actinide nuclei is given by equation (420), and where equations (41) and (42) give the following approximate expressions for the critical magnetic induction fields  $B_{aY}^F$  and  $B_{bY}^F$  required for  $\gamma$  ray catalyzed thermal neutron induced fission of the subactinide nuclei lobes<sup>26</sup>

$$B_{aY}^F = K_{\theta_z}^{B_Y}(a)(1 - \chi_a)^{1/2} \quad 0 \leq \chi_a \leq 1 \quad (424)$$

$$B_{bY}^F = K_{\theta_z}^{B_Y}(b)(1 - \chi_b)^{1/2} \quad 0 \leq \chi_b \leq 1 \quad (425)$$

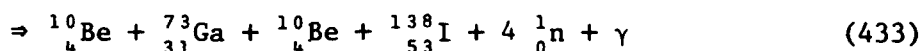
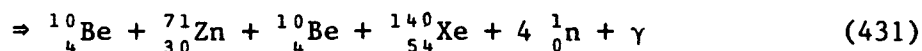
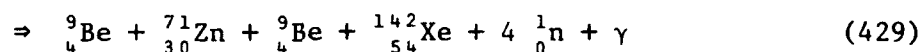
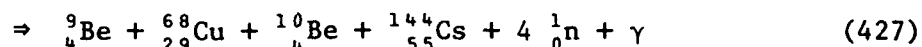
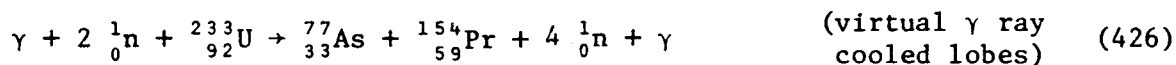
where  $\chi_a$  and  $\chi_b$  = fissility parameters of the subactinide component nuclei lobes a and b. The two subactinide nuclei lobes of the distorted binary fission suppressed actinide nucleus will undergo binary fission so that thermal neutron induced quaternary fission will be the dominant fission decay mode of a fissile actinide nucleus in a  $\gamma$  ray field that satisfies the conditions of equation (423). The characteristics of the electromagnetic field required for the  $\gamma$  ray catalyzed thermal neutron induced binary fission in the subactinide elements has been treated in the literature and will not be repeated here for the subactinide nuclear lobes a and b.<sup>26</sup>

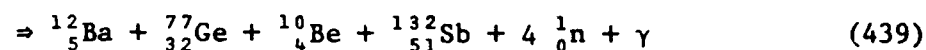
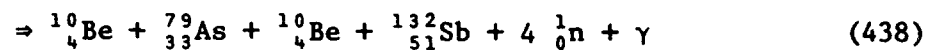
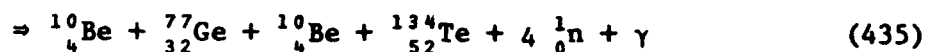
The fission product nuclei of the  $\gamma$  ray catalyzed quaternary fission process are relatively light nuclei which exhibit low level beta decays and are not harmful to the environment. Tables 1 through 3 give the relevant characteristics of the electromagnetic field of the  $\gamma$  rays that are required for the suppression of binary fission in the actinides and which are described by equations (392) and (420). The calculations involved in the preparation of Tables 1 through 3 are analogous to those given in Reference 26 except that now equations (392) and (420) determine the basic calculation of the electromagnetic field strength required for binary fission suppression in the actinides.

#### B. Examples of $\gamma$ Ray Catalyzed Thermal Neutron Induced Quaternary Fission of the Actinides.

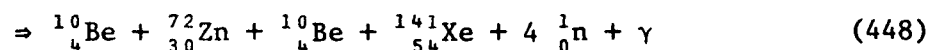
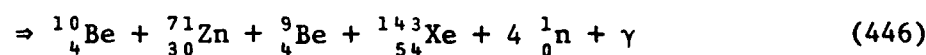
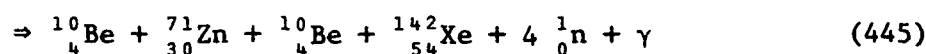
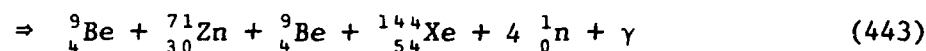
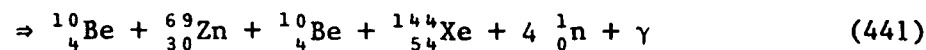
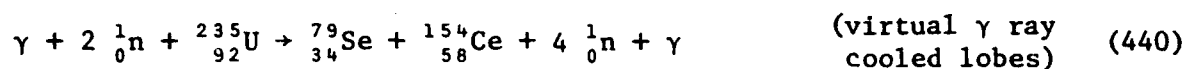
Consider now some typical examples of the quaternary fission of the fissile actinides by  $\gamma$  ray suppression of the binary fission of the parent fissile actinide nucleus and the  $\gamma$  ray catalyzation of thermal neutron induced binary fission of the subactinide nuclei lobes of the parent actinide nucleus. Typical reactions of this form require one neutron for each subactinide lobe, and therefore two incident neutrons are required for the quaternary fission of a fissile actinide nucleus. Therefore including absorption at least four fission product neutrons are required for each quaternary fission reaction in order to have the sustained fission reactions required for the operation of a nuclear reactor. In addition, a  $\gamma$  ray photon bath is required to suppress the binary fission process in the fissile actinides and to catalyze the binary fission of the subactinide lobes of the  $\gamma$  ray cooled actinide nuclei.

Typical  $\gamma$  ray catalyzed thermal neutron induced quaternary fission reactions for  $\gamma$  ray cooled fissile actinide nuclei such as  $^{233}\text{U}$ ,  $^{235}\text{U}$  and  $^{239}\text{Pu}$  will now be presented

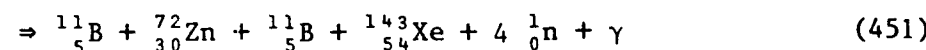
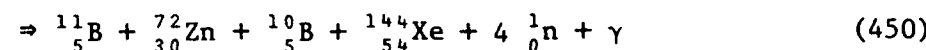
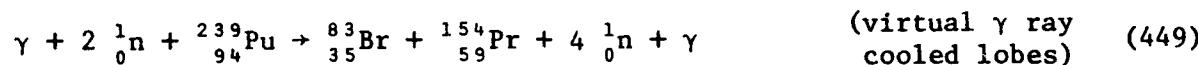


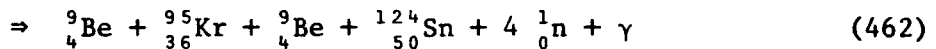
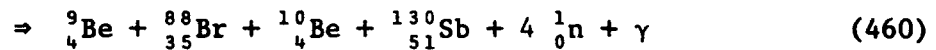
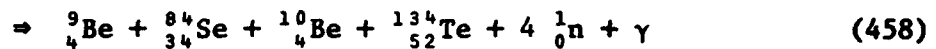
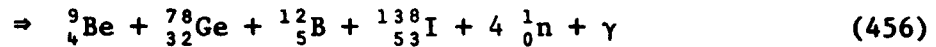
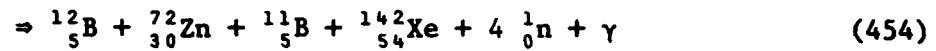
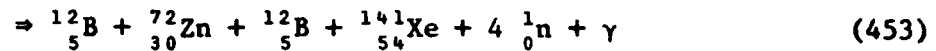


### ${}^{235}\text{U}$



### ${}^{239}\text{Pu}$





In this manner the dangerous fission products such as  ${}^{90}_{38}\text{Sr}$  can be eliminated from nuclear reactor wastes.

### C. $\gamma$ Ray Cooled Actinide Fission Reactors.

The fissile actinide elements can be used to power clean fission nuclear reactors if a sufficiently high flux of  $\gamma$  rays is employed to cool the actinides to the point where thermal neutron induced binary fission is suppressed and where  $\gamma$  ray catalyzed thermal neutron induced binary fission can occur in the subactinide component lobes of the  $\gamma$  ray cooled fissile actinides. The result is that  $\gamma$  ray catalyzed thermal neutron induced quaternary fission is the dominant fission mode for this type of nuclear reactor. Tables 1 through 3 give the characteristics of the  $\gamma$  ray fields required to suppress thermal neutron induced binary fission in selected actinide nuclei. These are only approximate results because the Bohr-Wheeler fission condition describes only the gross features of nuclear fission, and in fact Table 1 shows that an incorrect prediction is made for  ${}^{235}_{92}\text{U}$  by predicting that this nucleus is not fissile. Figures 6 and 7 present two  $\gamma$  ray cooled actinide quaternary fission reactor designs that can be used as clean fission nuclear reactors.

In order to insure clean fission any design of a  $\gamma$  ray cooled actinide quaternary fission reactor must insure that all fissile actinide nuclei, such as  $^{235}\text{U}$ , are cooled by  $\gamma$  rays when thermal neutrons are present, otherwise ordinary binary fission will occur and the nuclear reactor will run in an unclean mode. The problem here is that neutrons generally penetrate matter easily except for special materials such as graphite or water, while  $\gamma$  rays are readily absorbed in most materials although quartz is a reasonable transmitter. For  $\gamma$  ray catalyzed thermal neutron induced quaternary fission the  $\gamma$  rays and the thermal neutrons must coexist together in the regions of a fuel element such as  $^{235}\text{U}$ . The thermal neutrons must be prevented from penetrating into the  $^{235}\text{U}$  by themselves after the  $\gamma$  rays have been absorbed otherwise ordinary binary fission will occur. A possible way to accomplish this would be to have  $^{235}\text{U}$  fuel embedded in quartz which is relatively transparent to  $\gamma$  rays. The neutron source and the  $\gamma$  ray source should be in close proximity in order to insure that  $^{235}\text{U}$  is simultaneously under the influence of  $\gamma$  rays and thermal neutrons. Ideally the  $\gamma$  rays should originate from the decay of a radioactive source mixed with the  $^{235}\text{U}$  fuel, however because the energy of the  $\gamma$  rays needs to be in a range that corresponds to the giant dipole resonance frequency of actinide nuclei, 12-14 MeV, it is likely that an artificial source of  $\gamma$  rays will be required.<sup>26</sup>

#### D. Electric Dipole Sum Rule for $\gamma$ Ray Cooled Actinides.

This section considers the effect of  $\gamma$  ray cooling on the form of the electric dipole sum rule for actinide nuclei whose fissility parameters satisfy  $\chi \geq 1$ . Consider  $\gamma$  rays and thermal neutrons incident on actinide nuclei in which a  $\gamma$  ray cooling mode is induced with an associated depressed binary fission rate. The  $\gamma$  ray energies are in the range of 12-14 MeV for the excitation of the giant dipole resonance in the actinides.<sup>28-35,49</sup>

The conventional electric dipole sum rule for photonuclear reactions is written in the standard incoherent spacetime form as follows<sup>28-35,49</sup>

$$G_{\text{inc}} = \int \sigma \, d\epsilon = gZN/A \quad (463)$$

where  $\sigma$  = photonuclear reaction cross section for incoherent spacetime,  $Z$ ,  $N$  and  $A$  = atomic number, neutron number and atomic mass number of the target nucleus, and where<sup>48</sup>

$$g = 2\pi^2 e^2 \hbar / (m_{\text{av}} c) \sim 0.06 \text{ MeV b} \quad (464)$$

where the integral is taken over photon energies up to 30 MeV. The concept of the broken symmetry forms of the atomic number, neutron number and atomic mass number suggests that a complex number generalization of the photonuclear reaction sum rule should be written as

$$\bar{G} = \int \bar{\sigma} \, d\bar{\epsilon} = g\bar{z}\bar{n}/\bar{a} \quad (465)$$

where  $\bar{\sigma}$  = complex number photonuclear reaction cross section,  $\bar{G}$  = complex number integrated photonuclear cross section, and  $\bar{z}$ ,  $\bar{n}$  and  $\bar{a}$  = complex number atomic number, neutron number and atomic mass number respectively which are

given by equations (43) through (45). The complex numbers  $\bar{\sigma}$  and  $\bar{G}$  can be represented as

$$\bar{\sigma} = \sigma \exp(j\theta_{\sigma}) \quad \bar{G} = G \exp(j\theta_G) \quad (466)$$

which are complex numbers in an internal space. Equation (465) can be written for actinide nuclei using equation (46) as

$$G = gzn/a = gZN/A \sec \theta_z \sec \theta_n \sec^{-1} \theta_a \quad (467)$$

$$\theta_G = \theta_z + \theta_n - \theta_a \quad (468)$$

Equation (465) can also be written as

$$\bar{G} = \int_0^{\infty} \sigma \sec \beta_{\epsilon\epsilon} \exp[j(\theta_{\sigma} + \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\epsilon \quad (469)$$

$$= \int_0^{\pi/6} \sigma \epsilon \csc \beta_{\epsilon\epsilon} \exp[j(\theta_{\sigma} + \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\theta_{\epsilon} \quad (470)$$

where the complex number photon energy is written as

$$\bar{\epsilon} = \epsilon \exp(j\theta_{\epsilon}) \quad (471)$$

and where

$$\tan \beta_{\epsilon\epsilon} = \epsilon \partial \theta_{\epsilon} / \partial \epsilon \quad (472)$$

The component form of equations (469) and (470) are written as

$$G \cos \theta_G = \int_0^{\infty} \sigma \sec \beta_{\epsilon\epsilon} \cos(\theta_{\sigma} + \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (473)$$

$$= \int_0^{\pi/6} \sigma \epsilon \csc \beta_{\epsilon\epsilon} \cos(\theta_{\sigma} + \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (474)$$

$$G \sin \theta_G = \int_0^{\infty} \sigma \sec \beta_{\epsilon\epsilon} \sin(\theta_{\sigma} + \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (475)$$

$$= \int_0^{\pi/6} \sigma \epsilon \csc \beta_{\epsilon\epsilon} \sin(\theta_{\sigma} + \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (476)$$

which are generally valid equations. From equations (467) and (468) it follows that for actinide nuclei

$$G \cos \theta_G = gZN/A \sec \theta_z \sec \theta_n \sec^{-1} \theta_a \cos(\theta_z + \theta_n - \theta_a) \quad (477)$$

$$G \sin \theta_G = gZN/A \sec \theta_z \sec \theta_n \sec^{-1} \theta_a \sin(\theta_z + \theta_n - \theta_a) \quad (478)$$

Generally  $G \cos \theta_G > 0$  if  $Z > 0$ ,  $N > 0$  and  $A > 0$ .

The upper integration limit of  $\pi/6$  in equations (470), (474) and (476) arises from the conservation of momentum for the photon-nucleon interaction which can be written as

$$\bar{\epsilon}/c = h\bar{\nu}/c = m\bar{v} \quad (479)$$

where the complex number photon frequency  $\bar{\nu}$  and the complex number nucleon velocity  $\bar{v}$  are written as

$$\bar{\nu} = \nu \exp(j\theta_\nu) \quad \bar{v} = v \exp(j\theta_v) \quad (480)$$

Equations (479) and (480) can be written as

$$\epsilon/c = h\nu/c = mv \quad \theta_\epsilon = \theta_\nu = \theta_v = \theta_x - \theta_t \quad (481)$$

which are valid for the inelastic photonuclear reaction. For coherent spacetime  $\theta_x = \pi/3$ ,  $\theta_t = \pi/6$  and  $\theta_v = \pi/6$ , so that  $\theta_\epsilon = \pi/6$  for inelastic photonuclear reactions in coherent spacetime.

For nuclei whose fissility parameters are greater than unity,  $\chi \geq 1$ , and which have been cooled to the point of the suppression of thermal neutron induced binary fission by an external  $\gamma$  ray field, the critical internal phase angles of the atomic number, neutron number and atomic mass number associated with binary fission suppression are related by equation (384) as follows

$$\theta_a^S = 2\theta_z^S \sim 0.76\theta_n^S \quad \theta_n^S \sim 2.63\theta_z^S \quad (482)$$

and therefore equations (467), (468) and (482) give approximately

$$G \sim gZN/A (1 - \tan^2 \theta_z^S)(4 \cos^2 \theta_z^S - 3)^{-1} \quad (483)$$

$$\theta_G = 0 \quad (484)$$

From equations (475) and (484) it follows that for incipient fission spacetime is incoherent and

$$\theta_\sigma = 0 \quad \theta_\epsilon = 0 \quad \beta_{\epsilon\epsilon} = 0 \quad (485)$$

so that the photonuclear interaction for the case of fission must be scalar ( $\theta_\sigma = 0$ ) with incoherent photon interactions, and equations (473) and (483) through (485) become

$$G = \int_0^\infty \sigma d\epsilon \sim gZN/A (1 - \tan^2 \theta_z^S)(4 \cos^2 \theta_z^S - 3)^{-1} \quad (486)$$

However it has been shown in equation (392) that the approximate condition for  $\gamma$  ray induced binary fission suppression is given by

$$\tan \theta_z^S = [(\chi - 1)/\chi]^{1/2} \quad (487)$$

where  $\chi \geq 1$  is the fissility parameter. Combining equations (486) and (487) gives for incoherent photon interactions with the actinides and for  $\chi \sim 1$

$$\int_0^\infty \sigma \, d\epsilon \sim gZN/A \, \chi^{-1} (2\chi - 1)(3 - 2\chi)^{-1} \quad (488)$$

and therefore the fissility parameter enters into the electric dipole sum rule that includes catalyzed thermal neutron induced quaternary fission of the actinides which are immersed in a  $\gamma$  ray field that has suppressed binary fission. The evaluation of the electric dipole sum for the actinides is done only approximately and is valid only for small internal phase angles or equivalently equation (488) is valid only for  $\chi \sim 1$ .

**7. FINAL STATE ENERGY CONDITIONS FOR THE BINARY FISSION OF THE ACTINIDES IN AN EXTERNAL FIELD.** This section considers the binary fission of actinide nuclei in an electromagnetic or gravitational field that is not strong enough to suppress the binary fission process, so that according to equation (418) the condition that describes this is  $\theta_z < \theta_z^S$  or  $B_\gamma < B_\gamma^S$ . The fission products are subactinide nuclei and neutrons. A comparison is made between the initial and final energy states of an actinide nucleus that has undergone binary fission in the presence of an external electromagnetic or gravitational field. A fission reaction in which a nucleus  $(Z, A)$  has split into two nuclei  $(Z_1, A_1)$  and  $(Z_2, A_2)$  is written in the form<sup>39-44</sup>

$$(Z, A) \rightarrow (Z_1, A_1) + (Z_2, A_2) \quad (489)$$

Then the nucleus  $(Z_2, A_2)$  is assumed to eject a neutron

$$(Z_2, A_2) \rightarrow (Z_2, A_2 - 1) + (0, 1) \quad (490)$$

where in this notation  $(0, 1)$  is a single neutron. In this way the general process of nuclear fission can be represented by a nuclear transformation of the general form given in equation (489). In an external field the nuclei represented in equation (489) are also associated with complex atomic numbers, neutron numbers and atomic mass numbers that in analogy to equations (43) through (45) for actinide nuclei and equations (8) through (10) for subactinide nuclei are given by

$$\bar{z} = z \exp(j\theta_z) = Z \sec \theta_z \exp(j\theta_z) \quad (491)$$

$$\bar{n} = n \exp(j\theta_n) = N \sec \theta_n \exp(j\theta_n) \quad (492)$$

$$\bar{a} = a \exp(j\theta_a) = A \sec \theta_a \exp(j\theta_a) \quad (493)$$

$$\bar{z}_1 = z_1 \exp(j\theta_{z1}) = Z_1 \cos \theta_{z1} \exp(j\theta_{z1}) \quad (494)$$

$$\bar{n}_1 = n_1 \exp(j\theta_{n1}) = N_1 \cos \theta_{n1} \exp(j\theta_{n1}) \quad (495)$$

$$\bar{a}_1 = a_1 \exp(j\theta_{a1}) = A_1 \cos \theta_{a1} \exp(j\theta_{a1}) \quad (496)$$



$$\bar{z}_2 = z_2 \exp(j\theta_{z2}) = Z_2 \cos \theta_{z2} \exp(j\theta_{z2}) \quad (497)$$

$$\bar{n}_2 = n_2 \exp(j\theta_{n2}) = N_2 \cos \theta_{n2} \exp(j\theta_{n2}) \quad (498)$$

$$\bar{a}_2 = a_2 \exp(j\theta_{a2}) = A_2 \cos \theta_{a2} \exp(j\theta_{a2}) \quad (499)$$

The determination of the energy released during the fission reaction given in equation (489) requires that all of the nine internal phase angles that appear in equations (497) through (499) be determined, and the procedure for doing this will now be given.

A. Determination of the Internal Phase Angles of the Atomic Number, Neutron Number and Atomic Mass Number for the Initial and Final States of Fission of the Actinides in an Electromagnetic Field.

The nuclei involved in the fission reaction given by equation (489) are subject to the following scalar baryon number conservation equations<sup>26</sup>

$$A = Z + N \quad A_1 = Z_1 + N_1 \quad A_2 = Z_2 + N_2 \quad (500)$$

$$A = A_1 + A_2 \quad Z = Z_1 + Z_2 \quad N = N_1 + N_2 \quad (501)$$

In an external field the nuclei represented in equation (489) are subject to the following complex atomic number, neutron number and atomic mass number conservation equations similar to equation (54)<sup>26</sup>

$$\bar{a} + W = \bar{z} + \bar{n} \quad \bar{a}_1 + W_1 = \bar{z}_1 + \bar{n}_1 \quad \bar{a}_2 + W_2 = \bar{z}_2 + \bar{n}_2 \quad (502)$$

$$\bar{a} + W_a = \bar{a}_1 + \bar{a}_2 \quad \bar{z} + W_z = \bar{z}_1 + \bar{z}_2 \quad \bar{n} + W_n = \bar{n}_1 + \bar{n}_2 \quad (503)$$

Equations (502) and (503) show that all of the W's are not independent, and in fact they are subject to the following equation<sup>26</sup>

$$W - W_1 - W_2 = W_a - W_z - W_n \quad (504)$$

Equations (502) and (503) can be combined with equations (491) through (499) to yield the following twelve equations that are valid for the actinide elements and their subactinide fission product nuclei

$$A + W = Z + N \quad (505)$$

$$A \tan \theta_a = Z \tan \theta_z + N \tan \theta_n \quad (506)$$

$$A_1 \cos^2 \theta_{a1} + W_1 = Z_1 \cos^2 \theta_{z1} + N_1 \cos^2 \theta_{n1} \quad (507)$$

$$A_1 \cos \theta_{a1} \sin \theta_{a1} = Z_1 \cos \theta_{z1} \sin \theta_{z1} + N_1 \cos \theta_{n1} \sin \theta_{n1} \quad (508)$$

$$A_2 \cos^2 \theta_{a2} + W_2 = Z_2 \cos^2 \theta_{z2} + N_2 \cos^2 \theta_{n2} \quad (509)$$

$$A_2 \cos \theta_{a2} \sin \theta_{a2} = Z_2 \cos \theta_{z2} \sin \theta_{z2} + N_2 \cos \theta_{n2} \sin \theta_{n2} \quad (510)$$

$$A + W_a = A_1 \cos^2 \theta_{a1} + A_2 \cos^2 \theta_{a2} \quad (511)$$

$$A \tan \theta_a = A_1 \cos \theta_{a1} \sin \theta_{a1} + A_2 \cos \theta_{a2} \sin \theta_{a2} \quad (512)$$

$$Z + W_z = Z_1 \cos^2 \theta_{z1} + Z_2 \cos^2 \theta_{z2} \quad (513)$$

$$Z \tan \theta_z = Z_1 \cos \theta_{z1} \sin \theta_{z1} + Z_2 \cos \theta_{z2} \sin \theta_{z2} \quad (514)$$

$$N + W_n = N_1 \cos^2 \theta_{n1} + N_2 \cos^2 \theta_{n2} \quad (515)$$

$$N \tan \theta_n = N_1 \cos \theta_{n1} \sin \theta_{n1} + N_2 \cos \theta_{n2} \sin \theta_{n2} \quad (516)$$

where  $Z, N, A$ ;  $Z_1, N_1, A_1$  and  $Z_2, N_2, A_2$  are known quantities.

There are fifteen unknown quantities in the problem of the fission of an atomic nucleus in the presence of an electromagnetic field:

$$W, \theta_z, \theta_n, \theta_a \quad (517)$$

$$W_1, \theta_{z1}, \theta_{n1}, \theta_{a1} \quad (518)$$

$$W_2, \theta_{z2}, \theta_{n2}, \theta_{a2} \quad (519)$$

$$W_z, W_n, W_a \quad (520)$$

There are fifteen equations to determine these quantities and they are: the twelve equations (505) through (516), the two fission stability equations (356) and (357) which determine  $\theta_z$  and  $\theta_a$  in the forms of equations (365) and (366), and finally equation (504) which relates the various  $W$ -functions. The values of the  $W$ 's for actinide nuclei and their subactinide nuclei fission products are obtained from equations (505) through (516) to be

$$W = 0 \quad (521)$$

$$W_1 = -A_1/2[1 + (1 - 4f_{W1}^2)^{1/2}] + Z_1 \cos^2 \theta_{z1} + N_1 \cos^2 \theta_{n1} \quad (522)$$

$$W_2 = -A_2/2[1 + (1 - 4f_{W2}^2)^{1/2}] + Z_2 \cos^2 \theta_{z2} + N_2 \cos^2 \theta_{n2} \quad (523)$$

$$W_z = -Z + Z_1 \cos^2 \theta_{z1} + Z_2 \cos^2 \theta_{z2} \quad (524)$$

$$W_n = -N + N_1 \cos^2 \theta_{n1} + N_2 \cos^2 \theta_{n2} \quad (525)$$

$$W_a = -A + A_1 \cos^2 \theta_{a1} + A_2 \cos^2 \theta_{a2} \quad (526)$$

where

$$f_{W1} = A_1^{-1} (Z_1 \sin \theta_{z1} \cos \theta_{z1} + N_1 \sin \theta_{n1} \cos \theta_{n1}) \quad (527)$$

$$f_{W2} = A_2^{-1} (Z_2 \sin \theta_{z2} \cos \theta_{z2} + N_2 \sin \theta_{n2} \cos \theta_{n2}) \quad (528)$$

and where

$$\cos^2 \theta_{a1} = 1/2 [1 + (1 - 4f_{W1}^2)^{1/2}] \quad (529)$$

$$\cos^2 \theta_{a2} = 1/2 [1 + (1 - 4f_{W2}^2)^{1/2}] \quad (530)$$

$$\tan \theta_z = Z_1/Z \cos \theta_{z1} \sin \theta_{z1} + Z_2/Z \cos \theta_{z2} \sin \theta_{z2} \quad (531)$$

$$\tan \theta_n = N_1/N \cos \theta_{n1} \sin \theta_{n1} + N_2/N \cos \theta_{n2} \sin \theta_{n2} \quad (532)$$

$$\tan \theta_a = A_1/A \cos \theta_{a1} \sin \theta_{a1} + A_2/A \cos \theta_{a2} \sin \theta_{a2} \quad (533)$$

Equations (521) through (526) can be rewritten as

$$W = 0 \quad (534)$$

$$W_1 = -A_1 \cos^2 \theta_{a1} + Z_1 \cos^2 \theta_{z1} + N_1 \cos^2 \theta_{n1} \quad (535)$$

$$= Z_1 (\cos^2 \theta_{z1} - \cos^2 \theta_{a1}) + N_1 (\cos^2 \theta_{n1} - \cos^2 \theta_{a1})$$

$$W_2 = -A_2 \cos^2 \theta_{a2} + Z_2 \cos^2 \theta_{z2} + N_2 \cos^2 \theta_{n2} \quad (536)$$

$$= Z_2 (\cos^2 \theta_{z2} - \cos^2 \theta_{a2}) + N_2 (\cos^2 \theta_{n2} - \cos^2 \theta_{a2})$$

$$W_z = -Z_1 \sin^2 \theta_{z1} - Z_2 \sin^2 \theta_{z2} \quad (537)$$

$$W_n = -N_1 \sin^2 \theta_{n1} - N_2 \sin^2 \theta_{n2} \quad (538)$$

$$W_a = -A_1 \sin^2 \theta_{a1} - A_2 \sin^2 \theta_{a2} \quad (539)$$

Therefore for actinide nuclei

$$W = 0 \quad W_1 > 0 \quad W_2 > 0 \quad (540)$$

$$W_z < 0 \quad W_n < 0 \quad W_a < 0 \quad (541)$$

For the case of fission of actinide nuclei in an electromagnetic field only  $W = 0$  while the other  $W$  functions have nonzero values.

#### B. Energy Released from Nuclear Fission in an External Field.

The  $Q$  value of a nuclear reaction is a measure of the energy released in a nuclear fission process.<sup>29</sup> In this paper a complex number generalization of the standard definition of the  $Q$  value is given by<sup>26</sup>

$$\bar{Q}/c^2 = \bar{M}(A, Z) - \bar{M}(A_1, Z_1) - \bar{M}(A_2, Z_2) \quad (542)$$

where as in equation (294)<sup>26</sup>

$$\bar{M}(A, Z) = \bar{n}m_n + \bar{z}m_H - \bar{B}(A, Z)/c^2 \quad (543)$$

$$\bar{M}(A_1, Z_1) = \bar{n}_1m_n + \bar{z}_1m_H - \bar{B}(A_1, Z_1)/c^2 \quad (544)$$

$$\bar{M}(A_2, Z_2) = \bar{n}_2m_n + \bar{z}_2m_H - \bar{B}(A_2, Z_2)/c^2 \quad (545)$$

Then the  $\bar{Q}$  value can be written as

$$\begin{aligned} \bar{Q} = [(\bar{n} - \bar{n}_1 - \bar{n}_2)m_n + (\bar{z} - \bar{z}_1 - \bar{z}_2)m_H]c^2 \\ + \bar{B}(A_1, Z_1) + \bar{B}(A_2, Z_2) - \bar{B}(A, Z) \end{aligned} \quad (546)$$

Using equation (503) allows equation (546) to be written as

$$\bar{Q} = Q_1 + \bar{Q}_2 \quad (547)$$

where

$$Q_1 = - (W_n m_n + W_z m_H)c^2 \quad (548)$$

$$\bar{Q}_2 = \bar{B}(A_1, Z_1) + \bar{B}(A_2, Z_2) - \bar{B}(A, Z) \quad (549)$$

where  $W_z$  and  $W_n$  are given by equations (537) and (538) respectively.

Because  $W_n < 0$  and  $W_z < 0$  it follows that for actinide nuclei

$$Q_1 > 0 \quad (550)$$

The value of  $Q_1$  arises from the rest mass terms in equations (543) through (548). The actual rest mass is unchanged in a nuclear fission process because

$$Nm_n + Zm_H - (N_1m_n + Z_1m_H) - (N_2m_n + Z_2m_H) = 0 \quad (551)$$

which is always true because of the absolute validity of baryon number conservation which for the present case is written as

$$Z = Z_1 + Z_2 \quad N = N_1 + N_2 \quad (552)$$

A finite value of  $Q_1$  results from the special form of the conservation law of complex baryon numbers, which for the complex atomic number, neutron number and atomic mass number are given in equations (502) and (503). The expression for  $Q_1$  for actinide nuclei can be rewritten using equations (537), (538) and (548) as

$$Q_1/c^2 = m_n(N_1 \sin^2 \theta_{n1} + N_2 \sin^2 \theta_{n2}) + m_H(Z_1 \sin^2 \theta_{z1} + Z_2 \sin^2 \theta_{z2}) \quad (553)$$

In general  $Q_1 > 0$  for actinide nuclei. For zero value of the applied external field  $Q_1 = 0$  because all internal phase angles have zero values, and therefore  $W_n = 0$  and  $W_z = 0$ .

The value of  $\bar{Q}_2$  can be calculated by combining equations (236) and (549). This is easily done for symmetric fission and under the approximation

$$\theta_{z1} \sim \theta_{z2} \sim \theta_z \quad \theta_{n1} \sim \theta_{n2} \sim \theta_n \quad (554)$$

For symmetric fission equation (549) becomes

$$\bar{Q}_2 = 2\bar{B}(A/2, Z/2) - \bar{B}(A, Z) \quad (555)$$

Under these assumptions the value of  $\bar{Q}_2$  is given by the following complex number generalization of the standard scalar result<sup>29</sup>

$$\begin{aligned} \bar{Q}_2 &= (1 - 2^{1/3})\bar{\gamma}\bar{a}^{2/3} + (1 - 2^{-2/3})\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \\ &= -0.26\bar{\gamma}\bar{a}^{2/3} + 0.37\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \end{aligned} \quad (556)$$

The simple form in equation (556) results from the approximation given in equation (554). The value of  $\bar{Q}$  is then written as

$$\bar{Q} = Q_1 - 0.26\bar{\gamma}\bar{a}^{2/3} + 0.37\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \quad (557)$$

The measured value of  $\bar{Q}$  for the actinides is given by the real part of equation (557)

$$\begin{aligned} Q_m &= Q_1 - 0.26\gamma A^{2/3} \sec^{2/3} \theta_a \cos(\theta_\gamma + 2/3\theta_a) \\ &\quad + 0.37\delta Z^2 A^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \cos(\theta_\delta + 2\theta_z - 1/3\theta_a) \end{aligned} \quad (558)$$

Equation (558) can be compared to the conventionally calculated value of  $Q$  which is given by<sup>29</sup>

$$Q_c = -0.26\gamma A^{2/3} + 0.37\delta Z^2 A^{-1/3} \quad (559)$$

For the case of zero external field equation (558) reduces to equation (559).

A condition that determines the possibility of the final fission state to occur can be obtained from the Q value for the nuclear fission process.<sup>29</sup> The complex number generalization of this condition is

$$\bar{Q} \geq \bar{E}_c^* \quad (560)$$

where  $\bar{E}_c^*$  = complex number Coulomb potential energy of two spherical nuclei (Z/2, A/2) in geometrical contact. This Coulomb energy can be written as a simple complex number generalization of the standard scalar result<sup>29</sup>

$$\begin{aligned} \bar{E}_c^* &= 1/2 e^2 (\bar{z}/2)^2 / [\bar{b}(\bar{a}/2)^{1/3}] \\ &= 2^{1/3} (1/8) (5/3) \bar{\delta} \bar{z}^2 / \bar{a}^{1/3} = 0.262 \bar{\delta} \bar{z}^2 / \bar{a}^{1/3} \end{aligned} \quad (561)$$

where as before in equation (264)

$$\bar{\delta} = 3/5 e^2 / \bar{b} = 0.863 / \bar{b} = (0.863 / 1.523) \bar{k}_c \quad \text{MeV} \quad (562)$$

where  $\bar{b}$  = complex number radius parameter given by equation (226). For  $k_c = 1.35 \text{ fm}^{-1}$  as in equation (267) it follows that

$$\delta = 0.765 \text{ MeV} = 765 \text{ keV} \quad (563)$$

and  $\theta_\delta$  is given by equation (266).

Combining equations (557), (560) and (561) gives the final state fission energy condition as

$$-Q_1 + 0.26 \bar{\gamma} \bar{a}^{2/3} = 0.11 \bar{\delta} \bar{z}^2 / \bar{a}^{1/3} \quad (564)$$

This equation can be used instead of the incipient fission condition given in equation (355) to determine  $\theta_z$  and  $\theta_a$ . However because of the presence of the functions  $W_n$  and  $W_z$  the full set of thirteen equations (504) through (516) must be solved in conjunction with the two components of equation (564) which are for the actinide elements

$$-Q_1 + 0.26 \bar{\gamma} \bar{a}^{2/3} \sec^{2/3} \theta_a \cos(\theta_\gamma + 2/3 \theta_a) \quad (565)$$

$$= 0.11 \bar{\delta} \bar{z}^2 \bar{a}^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \cos(\theta_\delta + 2\theta_z - 1/3 \theta_a)$$

$$0.26 \bar{\gamma} \bar{a}^{2/3} \sec^{2/3} \theta_a \sin(\theta_\gamma + 2/3 \theta_a) \quad (566)$$

$$= 0.11 \bar{\delta} \bar{z}^2 \bar{a}^{-1/3} \sec^2 \theta_z \sec^{-1/3} \theta_a \sin(\theta_\delta + 2\theta_z - 1/3 \theta_a)$$

If  $W_n$  and  $W_z$  are neglected in equation (564) so that  $Q_1 = 0$ , then the final state fission condition can be written as

$$\bar{z}^2 / \bar{a} = \kappa' \bar{\gamma} / \bar{\delta} \quad \kappa' = 2.36 \quad (567)$$

Equation (567) is the same form as the incipient fission condition given in equation (355) and the same form of solution for  $\theta_z^S$  and  $\theta_a^S$  that appears in equations (365) and (392) can now be used to determine these phase angles for the final state fission condition. The condition for binary fission in an external field is then given by equation (383). Then the remaining thirteen equations (504) through (516) can be used to calculate the remaining thirteen functions listed in equations (517) through (520).

**8. CONCLUSION.** A method of obtaining clean fission nuclear power from actinide elements has been proposed that is based on the idea of using a  $\gamma$  ray field to suppress thermal neutron induced binary fission in the fissile actinides and of using the same  $\gamma$  ray field to catalyze thermal neutron induced binary fission in the two subactinide lobes of the distorted  $\gamma$  ray cooled actinide nuclei. In this way a thermal neutron induced quaternary fission process can occur in  $\gamma$  ray cooled actinides. The net result is that the fission product nuclei for quaternary fission of  $\gamma$  ray cooled actinides are smaller and less radioactive than the fission product nuclei of conventional nuclear fission reactions.

#### ACKNOWLEDGEMENT

I would especially like to thank Elizabeth K. Klein for typing and editing this paper.

#### REFERENCES

1. Pigford, T. H., "Environmental Aspects of Nuclear Energy Production," *Ann. Rev. Nucl. Part. Sci.*, Vol. 24, p. 515, 1974.
2. Roberts, L. E. J., "Radioactive Waste Management," *Ann. Rev. Nucl. Part. Sci.*, Vol. 40, p. 79, 1990.
3. Kathren, R. L., Radioactivity in the Environment, Harwood, New York, 1984.
4. Chapman, N. A. and McKinley, I. G., The Geological Disposal of Nuclear Waste, John Wiley, New York, 1987.
5. Berlin, R. E. and Stanton, C. C., Radioactive Waste Management, John Wiley, New York, 1989.
6. Tang, Y. S. and Saling, J. H., Radioactive Waste Management, Hemisphere Publishing Corp., New York, 1990.
7. Krauskopf, K. B., "Disposal of High-Level Nuclear Waste: Is it Possible?," *Science*, Vol. 249, p. 1231, 14 September 1990.
8. Bell, G. I. and Glasstone, S., Nuclear Reactor Theory, Krieger Publishing Co., Melbourne, FL, 1979.
9. Glasstone, S., Sourcebook on Atomic Energy, Krieger Publishing Co., Melbourne, FL, 1979.

10. Krupnick, A. J. and Portney, P. R., "Controlling Urban Air Pollution: A Benefit-Cost Assessment," *Science*, Vol. 252, p. 522, 26 Apr 1991.
11. Corcoran, E., "Cleaning Up Coal," *Scientific American*, p. 107, May 1991.
12. Stern, A. C., editor, Air Pollution, Vols. 1-5, Academic, New York, 1976.
13. Wark, K. and Warner, C. F., Air Pollution, Harper and Row, New York, 1981.
14. Campbell, I. M., Energy and the Atmosphere, John Wiley, New York, 1986.
15. Hodges, H., Technology in the Ancient World, Barnes & Noble, New York, 1992.
16. De Camp, L. S., The Ancient Engineers, Dorset Press, New York, 1990.
17. Landels, J. G., Engineering in the Ancient World, University of California Press, 1978.
18. Kirby, R. S., Withington, S., Darling, A. B. and Kilgour, F. G., Engineering in History, Dover, New York, 1990.
19. Howes, R. and Fainberg, A., The Energy Sourcebook, American Institute of Physics, New York, 1991.
20. Glasstone, S., Energy Deskbook, Van Nostrand Reinhold, New York, 1983.
21. Hiler, E. A. and Stout, B. A., editors, Biomass Energy, Texas A&M University Press, College Station, 1985.
22. Lindl, J. D., McCrory, R. L. and Campbell, E. M., "Progress Toward Ignition and Burn Propagation in Inertial Confinement Fusion," *Physics Today*, pg. 32, September 1992.
23. Hogan, W. J., Bangerter, R. and Kulcinski, G., "Energy from Inertial Fusion," *Physics Today*, pg. 42, September 1992.
24. Corday, J. G., Goldston, R. J. and Parker, R. R., "Progress Toward a Tokamak Fusion Reactor," *Physics Today*, p. 22, January 1992.
25. Callen, J. D., Carreras, B. A. and Stambaugh, R. D., "Stability and Transport Processes in Tokamak Plasmas," *Physics Today*, p. 34, January 1992.
26. Weiss, R. A., "Clean Fission Nuclear Reactors," Tenth Army Conference on Applied Mathematics and Computing, West Point, New York, ARO 93-1, p. 463, June 16-19, 1992.
27. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
28. Green, A. E. S., Nuclear Physics, McGraw-Hill, New York, 1955.
29. Evans, R. D., The Atomic Nucleus, McGraw-Hill, New York, 1955.



30. Eder, G., Nuclear Forces, MIT Press, Cambridge, 1968.
31. Elton, L. R. B., Introductory Nuclear Theory, Interscience, New York, 1959.
32. DeBenedetti, S., Nuclear Interactions, John Wiley, New York, 1964.
33. Blatt, J. M. and Weisskopf, V. F., Theoretical Nuclear Physics, John Wiley, New York, 1952.
34. Bethe, H. A. and Morrison, P., Elementary Nuclear Theory, John Wiley, New York, 1961.
35. Greiner, W., Nuclear Theory, Vols. 1-3, Elsevier, New York, 1976.
36. Weiss, R. A. and Cameron, A. G. W., "Equilibrium Theory of the Nuclear Symmetry Energy of Infinite Nuclear Matter," and "Equilibrium Theory of the Symmetry Energy of Finite Nuclei," Can. J. Phys., Vol. 47, P. 2171 and p. 2211, 1969.
37. Romer, A., editor, The Discovery of Radioactivity and Transmutation, Dover, New York, 1964.
38. Rasetti, F., Elements of Nuclear Physics, Prentice-Hall, New York, 1936.
39. Willets, L., Theories of Nuclear Fission, Oxford Univ. Press, New York, 1964.
40. Fong, P., Statistical Theory of Nuclear Fission, Gordon & Breach, New York, 1969.
41. Vandenbosch, R. and Huizenga, J. R., Nuclear Fission, Academic Press, New York, 1973.
42. Poenaru, D. N. and Ivascu, M. S., Particle Emission from Nuclei, Vols. 1-3, CRC Press, Boca Raton, 1988.
43. Wheeler, J. A., "Fission in 1939: The Puzzle and the Promise," Ann. Rev. Nucl. Part. Sci., Vol. 39, p. xiii, 1989.
44. Wagemans, C., The Nuclear Fission Process, CRC Press, Boca Raton, 1991.
45. Walker, T., Feng, P., Hoffmann, D. and Williamson, R. S., "Spin-Polarized Spontaneous-Force Atom Trap," Phys. Rev. Lett., Vol. 69, pg. 2168, 12 October 1992.
46. Kasevich, M. and Chu, S., "Laser Cooling below a Photon Recoil with Three-Level Atoms," Phys. Rev. Lett., Vol. 69, pg. 1741, 21 September 1992.
47. Chu, S., "Laser Trapping of Neutral Particles," Scientific American, pg. 71, February 1992.
48. Cohen-Tannoudji, C. N. and Phillips, W. D., "New Mechanisms for Laser Cooling," Physics Today, pg. 33, October 1990.
49. Hayward, E., "Photonuclear Reactions," article in Encyclopedia of Physics, Lerner, R. and Trigg, G., editors, VCH Publishers, New York, 1991.

Table 1. Electromagnetic Field Characteristics for the  $\gamma$  Ray  
Suppression of Binary Fission in the Actinides

Nucleus	$Z^2/A$	$\chi$	$\theta_z^S$ (degrees)	$B^S$ ( $10^{12}$ T)	$B_y^S$ ( $10^3$ T)
$^{262}\text{Unp}$	42.08	1.1608	20.41	8.77	10.59
$^{261}\text{Unq}$	41.44	1.1432	19.49	8.34	10.08
$^{260}\text{Lr}$	40.80	1.1256	18.47	7.87	9.52
$^{259}\text{No}$	40.17	1.1081	17.35	7.36	8.91
$^{258}\text{Md}$	39.54	1.0907	16.09	6.80	8.23
$^{257}\text{Fm}$	38.91	1.0734	14.65	6.16	7.46
$^{252}\text{Es}$	38.89	1.0729	14.61	6.14	7.47
$^{251}\text{Cf}$	38.26	1.0555	12.92	5.41	6.57
$^{247}\text{Bk}$	38.09	1.0508	12.40	5.18	6.32
$^{244}\text{Cm}$	37.77	1.0419	11.35	4.73	5.78
$^{247}\text{Cm}$	37.31	1.0293	9.58	3.98	4.85
$^{243}\text{Am}$	37.14	1.0245	8.79	3.65	4.46
$^{228}\text{U}$	37.12	1.0241	8.72	3.62	4.47
$^{239}\text{Pu}$	36.97	1.0199	7.95	3.29	4.04
$^{234}\text{Np}$	36.96	1.0196	7.89	3.27	4.02
$^{242}\text{Pu}$	36.51	1.0072	4.83	1.99	2.44
$^{237}\text{Np}$	36.49	1.0067	4.66	1.92	2.36
$^{233}\text{U}$	36.33	1.0021	2.62	1.08	1.33
$^{223}\text{Th}$	36.32	1.0020	2.56	1.05	1.31
$^{228}\text{Pa}$	36.32	1.0019	2.49	1.03	1.27
$^{234}\text{U}$	36.17	0.9978	0	0	0
$^{235}\text{U}$	36.02	0.9936	0	0	0
$^{231}\text{Pa}$	35.85	0.9889	0	0	0
$^{227}\text{Th}$	35.68	0.9844	0	0	0
$^{238}\text{U}$	35.56	0.9810	0	0	0
$^{228}\text{Th}$	35.53	0.9800	0	0	0
$^{232}\text{Th}$	34.91	0.9631	0	0	0
$^{227}\text{Ac}$	34.89	0.9626	0	0	0

**Table 2. Electromagnetic Field Characteristics for the  $\gamma$  Ray  
Suppression of Binary Fission in the Actinides**

Nucleus	$10^9 \zeta_r$	$K_{0z}^{B7}$ ( $10^4$ T)	k ( $10^{19}$ N/m)	$f_r$ ( $10^{21}$ Hz)	$\lambda_r$ (fm)
$^{262}\text{Unp}$	1.207	2.846	2.000	3.022	99.18
$^{261}\text{Unq}$	1.208	2.848	1.998	3.026	99.06
$^{260}\text{Lr}$	1.209	2.850	1.996	3.030	98.93
$^{259}\text{No}$	1.210	2.851	1.992	3.034	98.80
$^{258}\text{Md}$	1.211	2.854	1.990	3.038	98.67
$^{257}\text{Fm}$	1.211	2.855	1.988	3.042	98.55
$^{252}\text{Es}$	1.215	2.865	1.974	3.062	97.90
$^{251}\text{Cf}$	1.216	2.866	1.969	3.066	97.77
$^{247}\text{Bk}$	1.219	2.874	1.961	3.082	97.25
$^{244}\text{Cm}$	1.222	2.880	1.953	3.095	96.86
$^{247}\text{Cm}$	1.219	2.874	1.961	3.082	97.25
$^{243}\text{Am}$	1.223	2.882	1.951	3.099	96.72
$^{228}\text{U}$	1.236	2.913	1.910	3.166	94.69
$^{239}\text{Pu}$	1.226	2.890	1.940	3.116	96.19
$^{234}\text{Np}$	1.230	2.900	1.927	3.138	95.51
$^{242}\text{Pu}$	1.223	2.884	1.948	3.103	96.59
$^{237}\text{Np}$	1.228	2.894	1.934	3.125	95.92
$^{233}\text{U}$	1.231	2.902	1.923	3.143	95.38
$^{223}\text{Th}$	1.240	2.924	1.896	3.189	93.99
$^{228}\text{Pa}$	1.236	2.913	1.910	3.166	94.69
$^{234}\text{U}$	1.230	2.900	1.927	3.138	95.51
$^{235}\text{U}$	1.229	2.898	1.929	3.134	95.65
$^{231}\text{Pa}$	1.233	2.907	1.918	3.152	95.10
$^{227}\text{Th}$	1.237	2.915	1.907	3.170	94.55
$^{238}\text{U}$	1.227	2.892	1.937	3.121	96.06
$^{228}\text{Th}$	1.236	2.913	1.910	3.166	94.69
$^{232}\text{Th}$	1.232	2.904	1.921	3.147	95.24
$^{227}\text{Ac}$	1.237	2.915	1.907	3.170	94.55

**Table 3. Electromagnetic Field Characteristics for the  $\gamma$  Ray  
Suppression of Binary Fission in the Actinides**

Nucleus	$H_{\gamma}^S$ ( $10^9$ amp/m)	$E_{\gamma}^S$ ( $10^{12}$ volts/m)	$P_{\gamma}^S$ ( $10^{22}$ W/m <sup>2</sup> )	$\epsilon_{\gamma}^S$ (MeV)	$\Phi_{\gamma}^S$ ( $10^{34}$ m <sup>-2</sup> sec <sup>-1</sup> )	$n_{\gamma}^S$ ( $10^{25}$ m <sup>-3</sup> )
<sup>262</sup> Unp	8.43	3.18	1.34	12.50	0.61	2.03
<sup>261</sup> Unq	8.02	3.02	1.21	12.52	0.56	1.86
<sup>260</sup> Lr	7.58	2.86	1.08	12.53	0.50	1.67
<sup>259</sup> No	7.09	2.67	0.95	12.55	0.44	1.47
<sup>258</sup> Md	6.55	2.47	0.81	12.57	0.38	1.27
<sup>257</sup> Fm	5.94	2.24	0.67	12.58	0.31	1.03
<sup>252</sup> Es	5.94	2.24	0.67	12.67	0.31	1.03
<sup>251</sup> Cf	5.23	1.97	0.52	12.68	0.24	0.80
<sup>247</sup> Bk	5.03	1.90	0.48	12.75	0.22	0.73
<sup>244</sup> Cm	4.60	1.73	0.40	12.80	0.19	0.63
<sup>247</sup> Cm	3.86	1.46	0.28	12.75	0.13	0.43
<sup>243</sup> Am	3.55	1.34	0.24	12.82	0.11	0.37
<sup>228</sup> U	3.56	1.34	0.24	13.10	0.11	0.37
<sup>239</sup> Pu	3.21	1.21	0.19	12.89	0.09	0.30
<sup>234</sup> Np	3.20	1.21	0.19	12.98	0.09	0.30
<sup>242</sup> Pu	1.94	0.73	0.07	12.84	0.03	0.10
<sup>237</sup> Np	1.88	0.71	0.07	12.93	0.03	0.10
<sup>233</sup> U	1.06	0.40	0.02	13.00	0.01	0.03
<sup>223</sup> Th	1.04	0.39	0.02	13.19	0.01	0.03
<sup>228</sup> Pa	1.01	0.38	0.02	13.10	0.009	0.03
<sup>234</sup> U	0	0	0	12.98	0	0
<sup>235</sup> U	0	0	0	12.96	0	0
<sup>231</sup> Pa	0	0	0	13.04	0	0
<sup>227</sup> Th	0	0	0	13.11	0	0
<sup>238</sup> U	0	0	0	12.91	0	0
<sup>228</sup> Th	0	0	0	13.10	0	0
<sup>232</sup> Th	0	0	0	13.02	0	0
<sup>227</sup> Ac	0	0	0	13.11	0	0

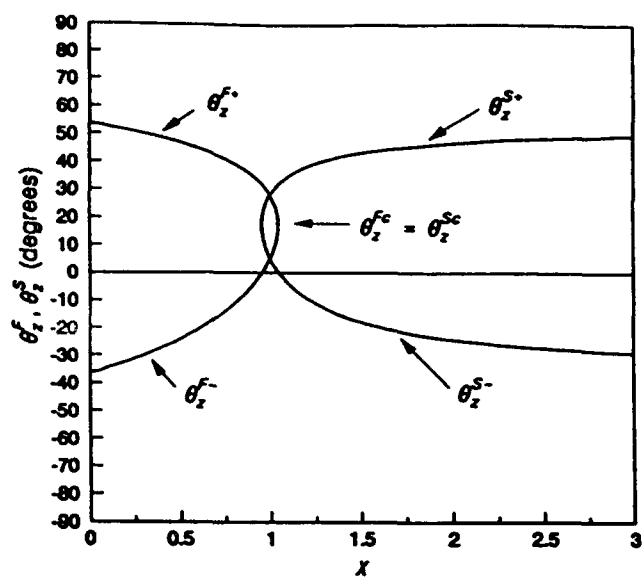


Figure 1. The fission angles  $\theta_1^f$  and  $\theta_1^s$  versus the fissility parameter  $\chi$  for the case  $\theta_1 = 0.4r = 22.9^\circ$  and  $\theta_s = 0.1r = 5.7^\circ$ .

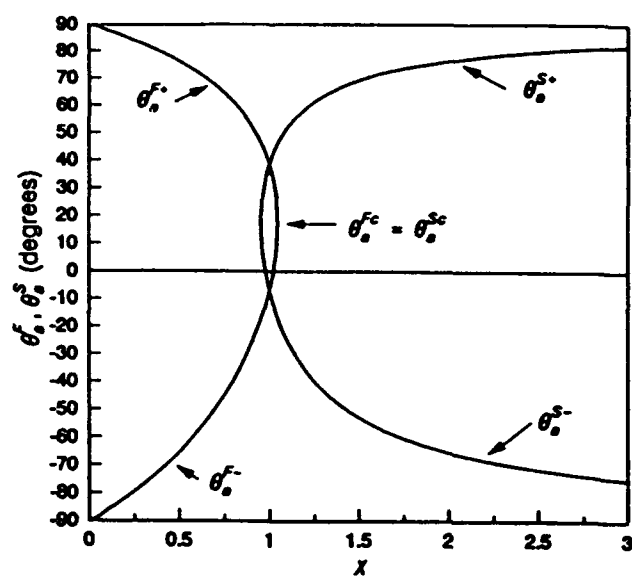


Figure 2. The fission angles  $\theta_2^f$  and  $\theta_2^s$  versus the fissility parameter  $\chi$  for the case  $\theta_1 = 0.4r = 22.9^\circ$  and  $\theta_s = 0.1r = 5.7^\circ$ .

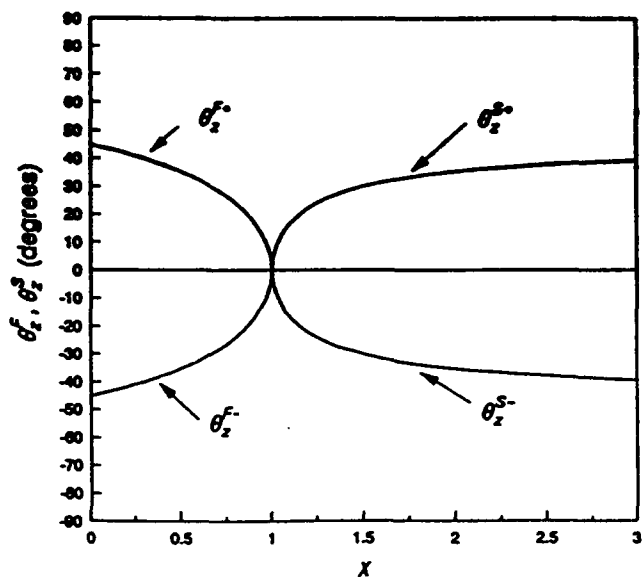


Figure 3. The fission angles  $\theta_1^f$  and  $\theta_1^s$  versus the fissility parameter  $\chi$  for the case  $\theta_1 = \theta_s$ .

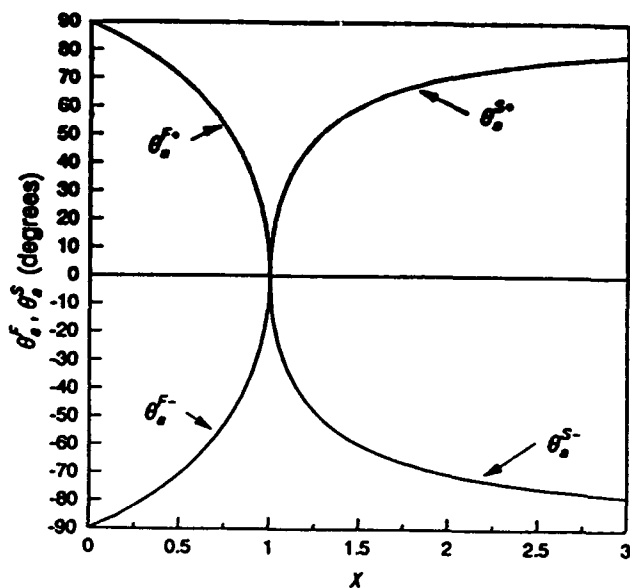


Figure 4. The fission angles  $\theta_2^f$  and  $\theta_2^s$  versus the fissility parameter  $\chi$  for the case  $\theta_1 = \theta_s$ .

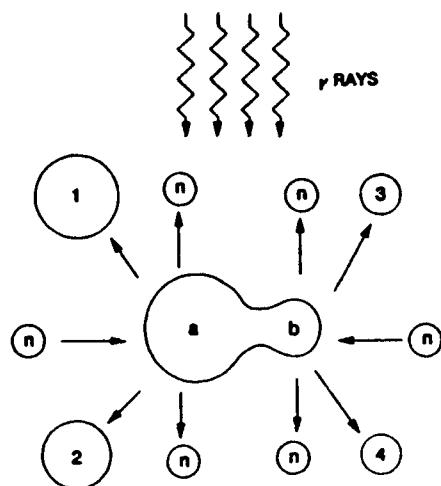


Figure 5. Schematic representation of the quaternary fission of a  $\gamma$  ray cooled actinide nucleus.

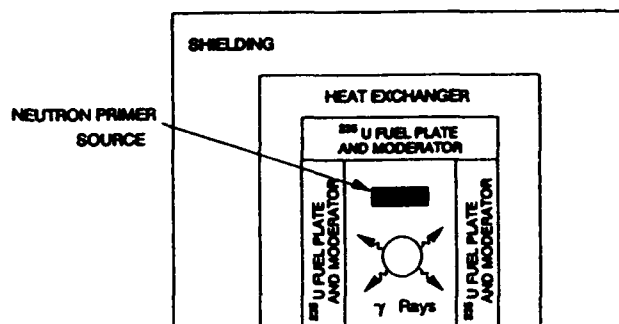


Figure 6. Design concept for a  $\gamma$  ray cooled actinide quaternary fission nuclear reactor.

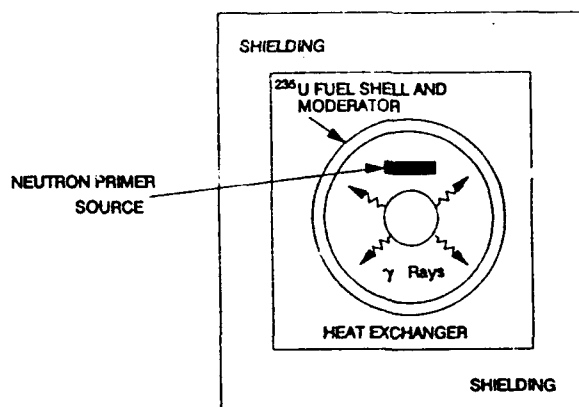


Figure 7. Design concept for a spherical  $\gamma$  ray cooled actinide quaternary fission nuclear reactor.

# Model Conversions of Uncertain Linear Systems Using a Scaling and Squaring Geometric Series Method

Leang S. Shieh†, Jingfong Gu†, and Jason S.H. Tsai‡

†Department of Electrical Engineering, University of Houston, Houston, Texas 77204-4793 USA.

‡Control Laboratory, Department of Electrical Engineering, National Cheng-Kung University, Tainan 70101, Taiwan, Republic of China.

## Abstract

This paper proposes a scaling and squaring geometric series method along with the inverse-geometric series method for finding the discrete-time (continuous-time) structured uncertain linear model from the continuous-time (discrete-time) structured uncertain linear systems. Above methods allow the use of well-developed theorems and algorithms in the discrete-time (continuous-time) domain to indirectly solve the continuous-time (discrete-time) domain problems. Moreover, these methods enhance the flexibility in modelling and control of a hybrid composite system. It has shown that the commonly used bilinear approximate model is a specific class of the proposed geometric series model.

## 1. Introduction

Most of the practical processes comprise of uncertain plants. The uncertainty about the plant arises from unmodelled dynamics, parameter variations, sensor noise, input signal level constraints, etc. Therefore, the real physical processes should be represented by a continuous-time and/or discrete-time uncertain framework. For digital simulation, parameter identification, hybrid control design and digital implementation of an uncertain linear system, it is essential to convert a continuous-time (discrete-time) uncertain linear system to an equivalent discrete-time (continuous-time) uncertain linear model. The model conversions of a nominal continuous-time (discrete-time) linear system to an equivalent nominal discrete-time (continuous-time) linear model have been reported in the literature [1,2,3]. However, the method for model conversions of uncertain linear state-space models has not yet been fully developed.

Ezzine and Johnson [4] used a classical perturbation method [5] to convert a continuous-time uncertain linear system to an equivalent discrete-time uncertain model but they did not solve the inverse of the problem. Shieh et al. [6] utilized the bilinear and inverse-bilinear transform method to carry out the model conversions. In this paper, we use a scaling and squaring geometric series method to perform the model conversions. We show that the proposed method significantly improves the accuracy of the existing models obtained by using the bilinear transform method [6].

## 2. Problem Formulation

Consider a structured continuous-time uncertain linear system

$$\dot{x}_c(t) = Ax_c(t) + Bu_c(t), \quad x_c(0) = x_{c0} \quad (1a)$$

$$y_c(t) = C_0x_c(t) \quad (1b)$$

where

$$A = A_0 + \Delta A = A_0 + \sum_{i=1}^{k_a} \Delta_{ai}A_i \quad (1c)$$

$$B = B_0 + \Delta B = B_0 + \sum_{i=1}^{k_b} \Delta_{bi}B_i \quad (1d)$$

$x_c(t) \in R^{n \times 1}$  is the state,  $u_c(t) \in R^{m \times 1}$  is the input,  $y_c(t) \in R^{p \times 1}$  is the output,  $(A_0, B_0, C_0)$  are nominal system matrices,  $(\Delta A, \Delta B)$  are perturbed uncertain matrices,  $(A_i, B_i)$  is the pair of known constant matrices,  $(\Delta_{ai}, \Delta_{bi})$  is the pair of uncertain scalar parameters. Without loss of generality, we can assume that  $|\Delta_{ai}| \leq 1$  for  $i = 1, 2, \dots, k_a$ , and  $|\Delta_{bi}| \leq 1$  for  $i = 1, 2, \dots, k_b$ .

The associated discrete-time uncertain linear model for (1) is

$$x_d(kT + T) = \hat{G}x_d(kT) + \hat{H}u_d(kT), \quad x_d(0) = x_{c0} \quad (2a)$$

$$y_d(kT) = C_0x_d(kT) \quad (2b)$$

where

$$\hat{G} = e^{(A_0 + \Delta A)T} \quad (2c)$$

$$\hat{H} = \int_0^T e^{(A_0 + \Delta A)\tau} (B_0 + \Delta B) d\tau = (\hat{G} - I_n)(A_0 + \Delta A)^{-1}(B_0 + \Delta B) \quad (2d)$$

$$u_d(t) = u_c(kT) \quad \text{for } kT \leq t < (k+1)T \quad (2e)$$



$T$  is the sampling period and  $I_n$  is an  $n \times n$  identity matrix. It is noted that  $e^{A_0 T}$  and  $e^{\Delta A T}$  are, in general, not commutative and  $\hat{G}$  and  $\hat{H}$  contain nonlinear uncertainty terms in  $\Delta A$  and  $(\Delta A, \Delta B)$ , respectively. In order to find the counterpart of the representation in (1) from (2), we linearize the respective  $\hat{G}$  in (2c) and  $\hat{H}$  in (2d) in the following manner.

$$\hat{G} \cong G_0 + \Delta G = G_0 + \sum_{i=1}^{h_a} \Delta_{a,i} G_i \quad (3a)$$

$$\hat{H} \cong H_0 + \Delta H = H_0 + \sum_{i=1}^{h_b} \Delta_{b,i} H_i \quad (3b)$$

where  $G_0 = e^{A_0 T}$  and  $H_0 = [G_0 - I_n] A_0^{-1} B_0 = \sum_{i=1}^{\infty} \frac{1}{i!} (A_0 T)^{i-1} B_0 T$ . All matrices in (2) and (3) have compatible terminologies and dimensions as their counterparts in (1). It is desired to find the pairs  $(G_0, H_0)$  and  $(\Delta G, \Delta H)$  in (3) from the pairs  $(A_0, B_0)$  and  $(\Delta A, \Delta B)$  in (1) such that the discretized state  $x_d(kT)$  in (2a) closely matches the state  $x_c(t)$  in (1a) at  $t = kT$  for a given piecewise-constant input  $u_d(t) = u_c(kT)$  for  $kT \leq t < (k+1)T$ .

Ezzine and Johnson[4] have shown that when the perturbation parameter  $|\Delta_{a,i}|$  in (1c) is sufficiently small and  $\Delta B = 0$ , then

$$\hat{G} = G_0 + \Delta \bar{G} + O(\Delta_{a,i}^2) \cong G_0 + \Delta \bar{G} \quad (4a)$$

$$\hat{H} = H_0 + \Delta \bar{H} + O(\Delta_{a,i}^2) \cong H_0 + \Delta \bar{H} \quad (4b)$$

where

$$\Delta \bar{G} = G_0 \int_0^T e^{-A_0 \tau} \Delta A e^{A_0 \tau} d\tau \quad (4c)$$

$$\Delta \bar{H} = \int_0^T \left( \int_0^T e^{A_0(t-\tau)} \Delta A e^{A_0 \tau} d\tau \right) B_0 dt. \quad (4d)$$

Note that  $O(\epsilon)$  denotes high-order terms in  $\epsilon$ , which can be neglected if  $\epsilon$  is sufficiently small. The determination of the exact  $\Delta \bar{G}$  in (4c) and  $\Delta \bar{H}$  in (4d) is not simple [5] because  $\lambda_i(-A_0) + \lambda_i(A_0) = 0$ , where  $\lambda_i(A_0)$  are the characteristic eigenvalues of  $A_0$ .

Shieh et al. [6] have shown that the uncertain system matrices are

$$\Delta G_b = \frac{1}{2} (G_0 - I_n) A_0^{-1} \Delta A (G_0 + I_n) \quad (5a)$$

$$\Delta H_b = (G_0 - I_n) A_0^{-1} \Delta B + \frac{1}{2} (G_0 - I_n) A_0^{-1} \Delta A H_0 \quad (5b)$$

and the counterparts of  $\Delta G (= \Delta G_b$  in (5a)) and  $\Delta H (= \Delta H_b$  in (5b)), defined as  $\Delta A_b$  and  $\Delta B_b$ , respectively, can be solved from respective (5a) and (5b) as

$$\Delta A_b = 2A_0(G_0 - I_n)^{-1}\Delta G(G_0 + I_n)^{-1} \quad (5c)$$

$$\Delta B_b = A_0(G_0 - I_n)^{-1}[\Delta H - \frac{1}{2}(G_0 - I_n)A_0^{-1}\Delta A_b H_0]. \quad (5d)$$

In their approach [6], the time-domain bilinear transformation with the form

$$e^{XT} \cong (I_n - \frac{1}{2}XT)^{-1}(I_n + \frac{1}{2}XT)$$

and the inverse-bilinear transformation with the form

$$(I_n - \frac{1}{2}XT)^{-1}(I_n + \frac{1}{2}XT) \cong e^{XT}$$

where  $X \in R^{n \times n}$ , were simultaneously utilized to derive (5).

### 3. The Geometric and Inverse-geometric Series Approximation Method

In this paper, we use a scaling and squaring geometric series method to linearize the respective  $\hat{G}$  and  $\hat{H}$  in (2). The geometric series method is described as follows.

The matrix-valued function of  $e^{XT}$  with  $X \in R^{n \times n}$  and a sampling period  $T$  can be approximated by a geometric series [7] as

$$\begin{aligned} \bar{G} &= e^{XT} = (e^{XT/m})^m = \left[ (e^{-\frac{1}{2}XT/m})^{-1} (e^{\frac{1}{2}XT/m}) \right]^m \\ &\cong (Q_j^{-1}P_j)^m \triangleq G_j \quad \text{for } j = 1, 2, \dots \end{aligned} \quad (6a)$$

where

$$Q_j = \left[ I_n - \frac{1}{2(j)(m)}XT \right] \left[ I_n + \sum_{i=1}^{j-1} \frac{(-1)^i(j-i)}{(2^i)(j)(i!)(m)^i} (XT)^i \right] \quad (6b)$$

$$P_j = \left[ I_n + \frac{1}{2(j)(m)}XT \right] \left[ I_n + \sum_{i=1}^{j-1} \frac{(j-i)}{(2^i)(j)(i!)(m)^i} (XT)^i \right] \quad (6c)$$

$$T < (2)(j)(m)/\|X\|. \quad (6d)$$

When  $j = 1, 2$  and  $3$ , we have

$$G_1 = \left\{ \left[ I_n - \frac{1}{2m}(XT) \right]^{-1} \left[ I_n + \frac{1}{2m}(XT) \right] \right\}^m \quad \text{for } T < \frac{2m}{\|X\|} \quad (7a)$$

$$G_2 = \left\{ \left[ I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^{-1} \left[ I_n + \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right] \right\}^m \quad (7b)$$

$$G_3 = \left\{ \left[ I_n - \frac{1}{2m}(XT) + \frac{7}{72m^2}(XT)^2 - \frac{1}{144m^3}(XT)^3 \right]^{-1} \times \right. \\ \left. \left[ I_n + \frac{1}{2m}(XT) + \frac{7}{72m^2}(XT)^2 + \frac{1}{144m^3}(XT)^3 \right] \right\}^m. \quad (7c)$$

The relationship between the sampling period  $T$  (defined as  $T_1$ ) in (7a) and the sampling period  $T$  (defined as  $T_2$ ) in (7b) can be determined by solving the following equation:

$$\left\{ \left( 1 - \frac{1}{2m} \lambda_s T_1 \right)^{-1} \left( 1 + \frac{1}{2m} \lambda_s T_1 \right) \right\}^m \\ = \left\{ \left[ 1 - \frac{1}{2m} \lambda_s T_2 + \frac{1}{16m^2} (\lambda_s T_2)^2 \right]^{-1} \left[ 1 + \frac{1}{2m} \lambda_s T_2 + \frac{1}{16m^2} (\lambda_s T_2)^2 \right] \right\}^m$$

where  $\lambda_s$  is the absolute value of the largest eigenvalue of  $X$ . Thus,

$$T_1 = \frac{1}{1 + \frac{1}{16m^2} (\lambda_s T_2)^2} T_2 < T_2. \quad (7d)$$

From (7d) we observe that a relatively larger sampling period  $T$  can be used for the approximation of  $e^{XT}$  if a more sophisticated approximate model is utilized.

It might be interesting to see the role of the scaling and squaring factor  $m$  in the approximation of  $e^{XT}$  in (6). When  $m = 1$ , the matrix  $G_1$  in (7a) can be represented by a geometric series as follows:

$$G_1 = (I_n - \frac{1}{2}XT)^{-1}(I_n + \frac{1}{2}XT) \quad \text{for } T < \frac{2}{\|X\|} \\ = I_n + XT + \frac{1}{2}(XT)^2 + \frac{1}{2^2(1)}(XT)^3 + \frac{1}{2^3(1)}(XT)^4 + \frac{1}{2^4(1)}(XT)^5 + \dots \quad (8a)$$

It is well-known [3] that the matrix  $G_1$  is the approximate model of  $e^{XT}$  obtained by using the Tustin method or bilinear transform method. When  $m = 2$ , (7a) becomes

$$G_1 = \left\{ \left( I_n - \frac{1}{2(2)}XT \right)^{-1} \left( I_n + \frac{1}{2(2)}XT \right) \right\}^2 \quad \text{for } T < \frac{2(2)}{\|X\|} \\ = I_n + XT + \frac{1}{2}(XT)^2 + \frac{1}{2^2(\frac{4}{3})}(XT)^3 + \frac{1}{2^3(2)}(XT)^4 + \frac{1}{2^4(\frac{16}{5})}(XT)^5 + \dots \quad (8b)$$

If  $m = 4$ , (7a) becomes

$$G_1 = \left\{ \left( I_n - \frac{1}{2(4)} XT \right)^{-1} \left( I_n + \frac{1}{2(4)} XT \right) \right\}^4 \quad \text{for } T < \frac{2(4)}{\|X\|}$$

$$= I_n + XT + \frac{1}{2}(XT)^2 + \frac{1}{2^2(\frac{16}{11})}(XT)^3 + \frac{1}{2^3(\frac{8}{3})}(XT)^4 + \frac{1}{2^4(\frac{256}{45})}(XT)^5 + \dots \quad (8c)$$

The exact Taylor series expansion of  $e^{XT}$  is

$$\bar{G} = e^{XT}$$

$$= I_n + XT + \frac{1}{2}(XT)^2 + \frac{1}{2^2(1.5)}(XT)^3 + \frac{1}{2^3(3)}(XT)^4 + \frac{1}{2^4(7.5)}(XT)^5 + \dots \quad (9)$$

Comparing (8) and (9) we observe that the first three terms in all equations are identical and each of the weighting factors for the other terms  $(XT)^j$  in (8) approaches the corresponding terms in (9) as the value of  $m$  increases. Therefore, (8c) is a better approximate model than those in (8a) and (8b).

The inverse-geometric series approximation of  $e^{XT}$  can be represented as follows.

From (7a) we have the inverse-geometric series approximations as

$$\left\{ \left[ I_n - \frac{1}{2m}(XT) \right]^{-1} \left[ I_n + \frac{1}{2m}(XT) \right] \right\}^m = G_1 \cong \bar{G} = e^{XT} \quad (10a)$$

$$\left[ I_n - \frac{1}{2m}(XT) \right]^{-m} = \left\{ \left( I_n - \frac{1}{2m}XT \right)^{-1} \left[ \left( I_n + \frac{1}{2m}XT \right) - \left( I_n - \frac{1}{2m}XT \right) \right] \left( \frac{XT}{m} \right)^{-1} \right\}^m$$

$$= \left[ (G_1^{1/m} - I_n) \left( \frac{XT}{m} \right)^{-1} \right]^m \cong \left[ (\bar{G}^{1/m} - I_n) \left( \frac{XT}{m} \right)^{-1} \right]^m \quad (10b)$$

$$\left[ I_n + \frac{1}{2m}(XT) \right]^m = \left[ I_n - \frac{1}{2m}(XT) \right]^m G_1 \cong \left[ I_n - \frac{1}{2m}(XT) \right]^m \bar{G}. \quad (10c)$$

The inverse-geometric series approximations in (10) can be justified by the same reasoning as the geometric series approximation in (7a). Note that when  $m = 1$  in (10), the inverse-geometric series approximations become the inverse-bilinear approximations.

When  $X = A_0 + \Delta A$ , the sampling period  $T$  for the sufficient condition  $\|XT\|/2m < 1$  in (7a) can be derived as follows:

Since

$$\frac{\|X\|T}{2m} = \frac{\|A_0 + \Delta A\|T}{2m} \leq \frac{(\|A_0\| + \|\Delta A\|)T}{2m} < 1 \quad (11a)$$

hence,

$$T < \frac{2m}{\|A_0\| + \|\Delta A\|}. \quad (11b)$$

Also, from (7b) we have the inverse-geometric series approximations as

$$\left\{ \left[ I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^{-1} \left[ I_n + \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right] \right\}^m = G_2 \cong \bar{G} = e^{XT} \quad (12a)$$

$$\begin{aligned} & \left[ I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^{-m} \\ &= \left\{ \left[ I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^{-1} \left[ \left( I_n + \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right) \right. \right. \\ & \quad \left. \left. - \left( I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right) \right] \left( \frac{XT}{m} \right)^{-1} \right\}^m \\ &= \left[ (G_2^{1/m} - I_n) \left( \frac{XT}{m} \right)^{-1} \right]^m \cong \left[ (\bar{G}^{1/m} - I_n) \left( \frac{XT}{m} \right)^{-1} \right]^m \end{aligned} \quad (12b)$$

$$\begin{aligned} & \left[ I_n + \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^m = \left[ I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^m G_2 \\ & \cong \left[ I_n - \frac{1}{2m}(XT) + \frac{1}{16m^2}(XT)^2 \right]^m \bar{G}. \end{aligned} \quad (12c)$$

In this paper, we carry out the model conversions by concentrating on the high-order bilinear approximate model in (7a) with  $m = 2$  which is identical to the approximate model in (7b) with  $m = 1$ . Then we compare the obtained model with the existing model in (5), which was obtained by a low-order bilinear approximation method.

#### 4. Model Conversions Using a High-Order Bilinear Approximation Method

Substituting  $X = A_0 + \Delta A$  into (6) and (7a) having  $m = 2$  gives

$$\begin{aligned} \hat{G}^{1/2} &\cong G_1^{1/2} \\ &= \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-1} \left[ I_n + \frac{(A_0 + \Delta A)T}{4} \right] \quad \text{for } T < \frac{4}{\|A_0\| + \|\Delta A\|} \end{aligned} \quad (13a)$$

Hence, we obtain

$$G_1^{1/2} - I_n = \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-1} \left[ \left( I_n + \frac{(A_0 + \Delta A)T}{4} \right) - \left( I_n - \frac{(A_0 + \Delta A)T}{4} \right) \right]$$

$$\begin{aligned}
&= \frac{T}{2} \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-1} (A_0 + \Delta A) \quad \text{for } T < \frac{4}{\|A_0\| + \|\Delta A\|} \\
G_1^{1/2} + I_n &= \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-1} \left[ \left( I_n + \frac{(A_0 + \Delta A)T}{4} \right) + \left( I_n - \frac{(A_0 + \Delta A)T}{4} \right) \right] \\
&= 2 \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-1} \quad \text{for } T < \frac{4}{\|A_0\| + \|\Delta A\|} \quad (13b)
\end{aligned}$$

and

$$\begin{aligned}
\hat{G} - I_n &\cong G_1 - I_n = (G_1^{1/2} - I_n)(G_1^{1/2} + I_n) \\
&= T \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-1} (A_0 + \Delta A). \quad (13c)
\end{aligned}$$

In the same manner, we substitute  $X = A_0$  into (6) and (7a) with  $m = 2$  yields

$$G_0^{1/2} - I_n \cong \frac{T}{2} \left( I_n - \frac{A_0 T}{4} \right)^{-1} A_0 \quad \text{for } T < \frac{4}{\|A_0\|} \quad (14a)$$

$$G_0^{1/2} + I_n \cong 2 \left( I_n - \frac{A_0 T}{4} \right)^{-1} \quad \text{for } T < \frac{4}{\|A_0\|} \quad (14b)$$

$$G_0 - I_n \cong T \left( I_n - \frac{A_0 T}{4} \right)^{-2} A_0 \quad (14c)$$

where  $G_0 = e^{A_0 T}$ . Also, Substituting  $X = A_0$ ,  $m = 2$  and  $\bar{G} = e^{A_0 T} = G_0$  in (10c) results in

$$\left( I_n + \frac{A_0 T}{4} \right) \cong \left( I_n - \frac{A_0 T}{4} \right) G_0^{1/2} \quad \text{for } T < \frac{4}{\|A_0\|}. \quad (14d)$$

The approximate model in (13c) can be further analyzed as follows.

$$\begin{aligned}
\hat{G} - I_n &\cong G_1 - I_n \\
&= \left[ I_n - \frac{(A_0 + \Delta A)T}{4} \right]^{-2} (A_0 + \Delta A)T \\
&= \left\{ \left[ I_n - \frac{A_0 T}{4} - \frac{\Delta A T}{4} \right] \left[ I_n - \frac{A_0 T}{4} - \frac{\Delta A T}{4} \right] \right\}^{-1} (A_0 + \Delta A)T \\
&= \left\{ \left[ I_n - \frac{A_0 T}{2} + \frac{(A_0 T)^2}{16} \right] \right. \\
&\quad \left. - \left[ \frac{\Delta A T}{2} - \frac{T^2}{16} (A_0 \Delta A + \Delta A A_0) \right] + O(T^2 \Delta A^2) \right\}^{-1} (A_0 + \Delta A)T \\
&\cong (W_0^{-1} - V_0)^{-1} (A_0 + \Delta A)T \quad (15a)
\end{aligned}$$

where

$$W_0 \triangleq \left[ I_n - \frac{A_0 T}{2} + \frac{(A_0 T)^2}{16} \right]^{-1} \quad (15b)$$

$$V_0 \triangleq \frac{1}{2} \Delta A T - \frac{T^2}{16} (A_0 \Delta A + \Delta A A_0) \quad (15c)$$

Note that the nonlinear uncertain term  $O(T^2 \Delta A^2)$  in (15a) has been neglected. Hence,

$$\begin{aligned} \hat{G} - I_n &\cong W_0 (I_n - V_0 W_0)^{-1} (A_0 + \Delta A) T \\ &= W_0 [I_n + V_0 W_0 + O((V_0 W_0)^2)] (A_0 + \Delta A) T \\ &\cong W_0 (I_n + V_0 W_0) (A_0 + \Delta A) T \quad \text{for } \|V_0 W_0\| < 1. \end{aligned} \quad (15d)$$

Note that the nonlinear uncertain term  $O((V_0 W_0)^2) = O(T^2 \Delta A^2)$  in (15d) is neglected. The sampling period  $T$  satisfies the convergent condition  $\|V_0 W_0\| < 1$  in (15d) as shown below:

Since  $\|V_0 W_0\| \leq \|V_0\| \|W_0\|$ , we can use the relationship in (12b) with  $X = A_0$ ,  $m = 1$  and  $\bar{G} = G_0$  to find  $\|W_0\|$ , i.e.,

$$\begin{aligned} \|W_0\| &\cong \|(G_0 - I_n)(A_0 T)^{-1}\| = \left\| \sum_{i=1}^{\infty} \frac{1}{i!} (A_0 T)^{i-1} \right\| \leq \sum_{i=1}^{\infty} \frac{1}{i!} \|A_0 T\|^{i-1} \\ &= 1 + \frac{\|A_0 T\|}{2!} + \frac{\|A_0 T\|^2}{3!} + \frac{\|A_0 T\|^3}{4!} + \frac{\|A_0 T\|^4}{(5)(4!)} + \frac{\|A_0 T\|^5}{(6)(5)(4!)} + \dots \\ &< 1 + \frac{\|A_0 T\|}{2!} + \frac{\|A_0 T\|^2}{3!} + \frac{\|A_0 T\|^3}{4!} \left[ 1 + \frac{\|A_0 T\|}{4} + \frac{\|A_0 T\|^2}{4^2} + \dots \right] \\ &= 1 + \frac{\|A_0 T\|}{2!} + \frac{\|A_0 T\|^2}{3!} + \frac{\|A_0 T\|^3}{4!} \left[ \frac{1}{1 - \frac{\|A_0 T\|}{4}} \right] \\ &= \frac{1 + \frac{\|A_0 T\|}{4} + \frac{\|A_0 T\|^2}{24}}{1 - \frac{\|A_0 T\|}{4}} \quad \text{for } T < \frac{4}{\|A_0\|} \end{aligned}$$

Also,

$$\|V_0\| \leq \frac{T}{2} \|\Delta A\| + \frac{T^2}{16} (2\|A_0\| \|\Delta A\|) = \frac{T}{2} \|\Delta A\| + \frac{T^2}{8} \|A_0\| \|\Delta A\|.$$

Therefore

$$\|V_0\| \|W_0\| < \frac{\frac{T^4}{192} \|A_0\|^3 \|\Delta A\| + \frac{5T^3}{96} \|A_0\|^2 \|\Delta A\| + \frac{T^2}{4} \|A_0\| \|\Delta A\| + \frac{T}{2} \|\Delta A\|}{1 - \frac{T}{4} \|A_0\|} < 1$$

i.e.

$$\frac{T^4}{192} \|A_0\|^3 \|\Delta A\| + \frac{5T^3}{96} \|A_0\|^2 \|\Delta A\| + \frac{T^2}{4} \|A_0\| \|\Delta A\| + \frac{T}{4} (\|A_0\| + 2\|\Delta A\|) - 1 < 0 \quad (15e)$$

Solving the inequality in (15e) for  $T$  results in the sufficient condition for the sampling period  $T$  such that  $\|V_0 W_0\| < 1$  and  $(I_n - V_0 W_0)^{-1} = I_n + V_0 W_0 + O((V_0 W_0)^2)$ . When the sampling period  $T$  in (15e) is sufficiently small, (15e) can be reduced to

$$\frac{T}{4} (\|A_0\| + 2\|\Delta A\|) < 1. \quad (15f)$$

As a result,

$$T < \frac{4}{\|A_0\| + 2\|\Delta A\|} < \frac{4}{\|A_0\| + \|\Delta A\|} < \frac{4}{\|A_0\|}. \quad (15g)$$

The results in (15e) and (15g) show that when the sampling period  $T$  satisfies the inequalities in (15e) and (15g), then

$$(I_n - V_0 W_0)^{-1} = I_n + V_0 W_0 + O(T^2 \Delta A^2) \cong I_n + V_0 W_0. \quad (15h)$$

Now we are ready to find the discrete-time uncertainties  $(\Delta G, \Delta H)$  in (3) from the continuous-time uncertainties  $(\Delta A, \Delta B)$  in (1) by neglecting any nonlinear uncertain terms  $O(T^2 \Delta A^2)$  and  $O(T^2 \Delta A \Delta B)$  in the following manner.

From (13a) and (14) we have

$$\begin{aligned} \hat{G} &\cong G_1 = \left\{ \left[ I_n - \frac{1}{4} (A_0 + \Delta A) T \right]^{-1} \left[ (I_n + \frac{1}{4} A_0 T) + \frac{1}{4} \Delta A T \right] \right\}^2 \\ &\cong \left\{ \left[ I_n - \frac{1}{4} (A_0 + \Delta A) T \right]^{-1} \left[ (I_n - \frac{1}{4} A_0 T) G_0^{1/2} + \frac{1}{4} \Delta A T \right] \right\}^2 \\ &= \left\{ \left[ I_n - \frac{1}{4} (A_0 + \Delta A) T \right]^{-1} \left[ \left( I_n - \frac{1}{4} (A_0 + \Delta A) T \right) G_0^{1/2} + \frac{1}{4} \Delta A T (G_0^{1/2} + I_n) \right] \right\}^2 \\ &= \left\{ G_0^{1/2} + \frac{T}{4} \left[ I_n - \frac{1}{4} (A_0 + \Delta A) T \right]^{-1} \Delta A (G_0^{1/2} + I_n) \right\}^2 \\ &= \left\{ G_0^{1/2} + \frac{T}{4} (I_n - \frac{1}{4} A_0 T)^{-1} \left[ I_n - \frac{1}{4} \Delta A T (I_n - \frac{1}{4} A_0 T)^{-1} \right]^{-1} \Delta A (G_0^{1/2} + I_n) \right\}^2 \end{aligned}$$



$$\begin{aligned}
&= \left\{ G_0^{1/2} + \frac{T}{4} (I_n - \frac{1}{4} A_0 T)^{-1} [I_n + \frac{T}{4} \Delta A (I_n - \frac{1}{4} A_0 T)^{-1} + O(T^2 \Delta A^2)] \Delta A (G_0^{1/2} + I_n) \right\}^2 \\
&\cong \left\{ G_0^{1/2} + \frac{T}{4} (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (G_0^{1/2} + I_n) \right. \\
&\quad \left. + (\frac{T}{4})^2 (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (G_0^{1/2} + I_n) \right\}^2 \\
&= \left\{ G_0^{1/2} + \frac{T}{4} (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (G_0^{1/2} + I_n) + O(T^2 \Delta A^2) \right\}^2 \\
&\cong \left\{ G_0^{1/2} + \frac{T}{4} (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (G_0^{1/2} + I_n) \right\}^2 \\
&= G_0 + \frac{T}{4} G_0^{1/2} (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (G_0^{1/2} + I_n) \\
&\quad + \frac{T}{4} (I_n - \frac{1}{4} A_0 T)^{-1} \Delta A (G_0^{1/2} + I_n) G_0^{1/2} + O(T^2 \Delta A^2) \\
&\cong G_0 + \frac{1}{2} G_0^{1/2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0^{1/2} + I_n) + \frac{1}{2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0^{1/2} + I_n) G_0^{1/2} \\
&= G_0 + \Delta G_1 \tag{16a}
\end{aligned}$$

where

$$G_0 = e^{A_0 T} \tag{16b}$$

$$G_0^{1/2} = e^{A_0 T/2} \tag{16c}$$

$$\begin{aligned}
\Delta G_1 &= \frac{1}{2} G_0^{1/2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0^{1/2} + I_n) \\
&\quad + \frac{1}{2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0^{1/2} + I_n) G_0^{1/2}. \tag{16d}
\end{aligned}$$

In order to derive the relationship between the obtained uncertain system matrix  $\Delta G_1$  in (16d) and the uncertain system matrix [6] as shown in (5a), we modify the representation in (16d) as

$$\begin{aligned}
\Delta G_1 &= \frac{1}{2} (G_0^{1/2} + I_n) (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0 + I_n - G_0 + G_0^{1/2}) \\
&\quad - \frac{1}{2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0^{1/2} + I_n) + \frac{1}{2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0^{1/2} + I_n) G_0^{1/2} \\
&= \Delta G_b + \frac{1}{2} (G_0^{1/2} - I_n) A_0^{-1} \Delta A (G_0 - I_n) \\
&\quad - \frac{1}{2} (G_0 - I_n) A_0^{-1} \Delta A (G_0^{1/2} - I_n) G_0^{1/2} \tag{17a}
\end{aligned}$$

where the uncertainty

$$\Delta G_b = \frac{1}{2}(G_0 - I_n)A_0^{-1}\Delta A(G_0 + I_n) \quad (17b)$$

is the discrete-time bilinear and inverse-bilinear uncertain system matrix [6] as shown in (5a). Also,

$$\hat{H} = \int_0^T e^{(A_0 + \Delta A)t} B dt = (\hat{G} - I_n)(A_0 + \Delta A)^{-1}(B_0 + \Delta B). \quad (18a)$$

Utilizing the result in (15d) gives

$$\begin{aligned} \hat{H} &\cong W_0(I_n + V_0 W_0)(B_0 + \Delta B)T \quad \text{for } \|V_0 W_0\| < 1 \\ &= W_0 B_0 T + W_0 V_0 W_0 B_0 T + W_0 \Delta B T + W_0 V_0 W_0 \Delta B T \\ &\cong H_0 + W_0 V_0 W_0 B_0 T + W_0 \Delta B T + O(T^2 \Delta A \Delta B) \\ &\cong H_0 + \Delta H_1 \end{aligned} \quad (18b)$$

where

$$\begin{aligned} G_0 &= e^{A_0 T} \\ W_0 &= \left[ I_n - \frac{1}{2} A_0 T + \frac{1}{16} (A_0 T)^2 \right]^{-1} \cong (G_0 - I_n)(A_0 T)^{-1} \\ V_0 &= \frac{1}{2} \Delta A_0 T - \frac{T^2}{16} (A_0 \Delta A + \Delta A A_0) \\ H_0 &= (G_0 - I_n) A_0^{-1} B_0 \cong W_0 B_0 T \\ \Delta H_1 &= W_0 V_0 W_0 B_0 T + W_0 \Delta B T \\ &= (G_0 - I_n) A_0^{-1} \left[ \frac{1}{2} \Delta A - \frac{T}{16} (A_0 \Delta A + \Delta A A_0) \right] H_0 + (G_0 - I_n) A_0^{-1} \Delta B \\ &= \Delta H_b - \frac{T}{16} (G_0 - I_n) A_0^{-1} (A_0 \Delta A + \Delta A A_0) H_0 \end{aligned} \quad (18c)$$

where the uncertainty

$$\Delta H_b = (G_0 - I_n) A_0^{-1} \Delta B + \frac{1}{2} (G_0 - I_n) A_0^{-1} \Delta A H_0 \quad (18d)$$

is the discrete-time bilinear and inverse-bilinear uncertain input matrix [6] as shown in (5b).

The counterpart of  $\Delta G$  (i.e.  $\Delta A$ ) can be determined from (16d) by solving the following Lyapunov equation [5]:

$$G_0^{1/2} \Delta A + \Delta A G_0^{1/2} = 2A_0(G_0^{1/2} - I_n)^{-1} \Delta G (G_0^{1/2} + I_n)^{-1} \quad (19a)$$

where

$$A_0 = \frac{1}{T} \ln(G_0). \quad (19b)$$

A computation method is available in [2] for computing the matrix logarithm function in (19b).

The counterpart of  $\Delta H$  (i.e.  $\Delta B$ ) can be directly obtained from (18c) as

$$\Delta B = A_0(G_0 - I_n)^{-1} \left\{ \Delta H - (G_0 - I_n)A_0^{-1} \left[ \frac{1}{2} \Delta A - \frac{T}{16} (A_0 \Delta A + \Delta A A_0) \right] H_0 \right\} \quad (20a)$$

where

$$A_0 = \frac{1}{T} \ln(G_0) \quad (20b)$$

$$B_0 = A_0(G_0 - I_n)^{-1} H_0. \quad (20c)$$

For the comparison of the aforementioned models in (5), (16), and (18) with the commonly used Taylor-series approximate model, we derive the Taylor-series approximate model as follows:

$$\begin{aligned} \hat{G} &= e^{(A_0 + \Delta A)T} \cong I_n + (A_0 + \Delta A)T + \frac{1}{2}(A_0 + \Delta A)^2 T^2 \\ &= (I_n + A_0 T + \frac{1}{2} A_0^2 T^2) + \Delta A (T I_n + \frac{T^2}{2} A_0) + (\frac{T^2}{2} A_0) \Delta A + O(T^2 \Delta A^2) \\ &\cong G_0 + \Delta G_t \end{aligned} \quad (21a)$$

where

$$G_0 = e^{A_0 T} \quad (21b)$$

$$\Delta G_t = \Delta A (T I_n + \frac{T^2}{2} A_0) + (\frac{T^2}{2} A_0) \Delta A. \quad (21c)$$

The corresponding discrete-time input matrix becomes

$$\begin{aligned}\hat{H} &= (\hat{G} - I_n)(A_0 + \Delta A)^{-1}(B_0 + \Delta B) \\ &= T(I_n + \frac{1}{2}A_0T)B_0 + \frac{T^2}{2}\Delta AB_0 + (TI_n + \frac{T^2}{2}A_0)\Delta B \\ &\cong H_0 + \Delta H_t\end{aligned}\quad (22a)$$

where

$$H_0 = (G_0 - I_n)A_0^{-1}B_0 \quad (22b)$$

$$\Delta H_t = \Delta A(\frac{T^2}{2}B_0) + (TI_n + \frac{T^2}{2}A_0)\Delta B. \quad (22c)$$

Also, the counterpart of  $\Delta G_t$  (i.e.  $\Delta A_t$ ) can be determined from (21c) by solving the following Lyapunov equation:

$$\Delta A_t(TI_n + \frac{T^2}{2}A_0) + (\frac{T^2}{2}A_0)\Delta A_t = \Delta G \quad (23a)$$

where  $A_0 = \frac{1}{T} \ln(G_0)$ . The counterpart of  $\Delta H_t$  (i.e.  $\Delta B_t$ ) can be directly solved from (22c) as

$$\Delta B_t = \frac{1}{T}(I_n + \frac{A_0T}{2})^{-1} \left[ \Delta H - \Delta A_t(\frac{T^2}{2}B_0) \right] \quad (23b)$$

where  $B_0 = A_0(G_0 - I_n)^{-1}H_0$ .

## 5. Illustrative Example

The unstable dynamics of a helicopter in a vertical plane for an airspeed range of 60 ~ 170 knots are given in [8, 9]. The nominal and uncertain system matrices are

$$A_0 = \begin{bmatrix} -0.0366 & 0.0271 & 0.0188 & -0.4555 \\ 0.0482 & -1.010 & 0.0024 & -4.0208 \\ 0.1002 & 0.2855 & -0.707 & 1.3229 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0.4422 & 0.1761 \\ 3.0447 & -7.5922 \\ -5.52 & 4.99 \\ 0 & 0 \end{bmatrix}$$

$$\Delta A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \bar{\Delta}_{a1} & 0 & \bar{\Delta}_{a2} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Delta B = \begin{bmatrix} 0 & 0 \\ \bar{\Delta}_{b1} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

where  $|\bar{\Delta}_{a1}| \leq 0.2192$ ,  $|\bar{\Delta}_{a2}| \leq 1.2031$  and  $|\bar{\Delta}_{b1}| \leq 2.0673$ .

The above perturbation parameters can be normalized as  $\Delta_{a1} = \Delta_{a2} = \Delta_{b1} = \Delta_b = \pm 1$ . Thus  $\Delta A = \Delta_b A_1 + \Delta_b A_2$  and  $\Delta B = \Delta_b B_1 + \Delta_b B_2$ , where  $B_2 = O_{4 \times 2}$  and

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.2192 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.2031 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 \\ 2.0673 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Let the sampling period

$$T = 0.2 < \frac{4}{\|A_0\| + \|\Delta A\|}.$$

The exact uncertain system matrix of the discrete-time model for  $|\Delta_b| = 1 \triangleq \epsilon$  are

$$\Delta G = e^{(A_0 + \epsilon A_1 + \epsilon A_2)T} - e^{A_0 T} = \epsilon \begin{bmatrix} 0.00000 & -0.00005 & -0.00001 & -0.00028 \\ -0.00000 & -0.00103 & -0.00028 & -0.00571 \\ 0.00034 & 0.03796 & 0.02114 & 0.21400 \\ 0.00002 & 0.00397 & 0.00145 & 0.02216 \end{bmatrix}$$

$$\begin{aligned} \Delta H &= (e^{(A_0 + \epsilon A_1 + \epsilon A_2)T} - I_4)(A_0 + \epsilon A_1 + \epsilon A_2)^{-1}(B_0 + \epsilon B_1) - (e^{A_0 T} - I_4)A_0^{-1}B_0 \\ &= \epsilon \begin{bmatrix} 0.00104 & 0.00001 \\ 0.37403 & 0.00034 \\ 0.02285 & -0.00036 \\ 0.00170 & -0.00168 \end{bmatrix} \end{aligned}$$

The uncertain matrices  $(\Delta G_1, \Delta H_1)$ , obtained by using the high-order bilinear and inverse-bilinear approximation method with  $m = 2$ , can be computed from (16d) and (18c), respectively, as

$$\Delta G_1 = \epsilon \begin{bmatrix} 0.00000 & -0.00005 & -0.00001 & -0.00027 \\ -0.00001 & -0.00101 & -0.00034 & -0.00566 \\ 0.00035 & 0.03774 & 0.02075 & 0.21269 \\ 0.00003 & 0.00393 & 0.00159 & 0.02201 \end{bmatrix}, \quad \Delta H_1 = \epsilon \begin{bmatrix} 0.00104 & 0.00001 \\ 0.37416 & 0.00038 \\ 0.01365 & -0.02166 \\ 0.00103 & -0.00172 \end{bmatrix}$$

The associated errors are  $\|\Delta G - \Delta G_1\|/\|\Delta G\| = 0.00633$ , and  $\|\Delta H - \Delta H_1\|/\|\Delta H\| = 0.02484$ .

Also, the uncertain matrices  $(\Delta G_b, \Delta H_b)$ , obtained by using the low-order bilinear and inverse-bilinear approximation method [6], can be computed from (5a) and (5b), respectively, as

$$\Delta G_b = \epsilon \begin{bmatrix} 0.00000 & -0.00005 & -0.00002 & -0.00027 \\ -0.00001 & -0.00099 & -0.00052 & -0.00556 \\ 0.00039 & 0.03801 & 0.01980 & 0.21384 \\ 0.00004 & 0.00387 & 0.00202 & 0.02180 \end{bmatrix}, \quad \Delta H_b = \epsilon \begin{bmatrix} 0.00104 & 0.00002 \\ 0.37424 & 0.00048 \\ 0.01067 & -0.01831 \\ 0.00073 & -0.00187 \end{bmatrix}$$

The associated errors are  $\|\Delta G - \Delta G_b\|/\|\Delta G\| = 0.00673$ , and  $\|\Delta H - \Delta H_b\|/\|\Delta H\| = 0.03477$ .

Moreover, the uncertain matrices  $(\Delta G_t, \Delta H_t)$ , obtained by using the Taylor series approximation method, can be computed from (21c) and (22c), respectively, as

$$\Delta G_t = \epsilon \begin{bmatrix} 0.00000 & 0.00008 & 0.00000 & 0.00045 \\ 0.00000 & 0.00001 & 0.00000 & 0.00006 \\ 0.00021 & 0.03631 & 0.02407 & 0.20598 \\ 0.00000 & 0.00438 & 0.00000 & 0.02406 \end{bmatrix} \quad \Delta H_t = \epsilon \begin{bmatrix} 0.00112 & 0.00000 \\ 0.37170 & 0.00000 \\ 0.02515 & -0.03328 \\ 0.00000 & 0.00000 \end{bmatrix}$$

The associated errors are  $\|\Delta G - \Delta G_t\|/\|\Delta G\| = 0.04815$ , and  $\|\Delta H - \Delta H_t\|/\|\Delta H\| = 0.02901$ .

The proposed approximate models are quite satisfactory.

In order to demonstrate the sensitivity of the proposed method to the sampling period ( $T \leq \frac{2m}{\|A_0\| + \|\Delta A\|}$ ) and the uncertain matrices  $(\Delta A, \Delta B)$ , the comparison of relative errors of the exact matrices  $(G, H)$  and any proposed approximate model  $(G_z, H_z)$ , i.e.,  $(\|G - G_z\|)/\|G\|$  and  $(\|H - H_z\|)/\|H\|$ , is presented in Fig.1 and Fig.2 with  $\Delta_h = 1$  and in Fig.3 and Fig.4 with  $\Delta_h = 0.1$ . It is observed that when the uncertainties  $(\Delta A, \Delta B)$  are sufficiently small, the obtained uncertain linear model using a high-order geometric series approximate model with a relatively large sampling period often gives a better result than those of lower-order geometric series approximate models with the same sampling period. On the other hand, for relatively large uncertainties  $(\Delta A, \Delta B)$ , the above observation may be not true due to the fact that the neglected nonlinear terms,  $O(T^2 \Delta A^2) = O[(\frac{2m}{\|A_0\| + \|\Delta A\|})^2 \Delta A^2]$  and  $O(T^2 \Delta A \Delta B) = O[(\frac{2m}{\|A_0\| + \|\Delta A\|})^2 \Delta A \Delta B]$ , are not sufficiently small.

## 6. Conclusion

A geometric series method along with the inverse-geometric series method has been proposed to find the discrete-time (continuous-time) structured uncertain linear models from the continuous-time (discrete-time) structured uncertain linear systems. This allows the use of well-developed theorems and algorithms in the discrete-time (continuous-time) domain to solve the continuous-time (discrete-time) domain problems indirectly. It has been shown that uncertain system parameters are propagated into the uncertain input matrix. Moreover, from Fig.3 and Fig.4 it has been shown that for a relatively large sampling

period and relatively small uncertain parameters ( $\epsilon < 1$ ), the models obtained by using a high-order geometric series approximation method are much better than those obtained by using the low-order geometric series approximation method. A numerical example is presented to illustrate the proposed procedures and to demonstrate the effectiveness of the proposed method.

## 7. Acknowledgment

This work was supported in part by the U.S. Army Research Office under contract DAAL-03-91-G0106, the NASA-Johnson Space Center under grant NAG 9-380 and the National Science Council of the Republic of China under grant NSC-81-0404-E-006-572.

## Reference

- [1] C. Moler and C. VanLoan, "Nineteen dubious ways to compute the exponential of a matrix", SIAM Rev., Vol. 20, PP. 801-836, 1978
- [2] L.S. Shieh, J.S.H. Tsai and S.R. Lian, "Determining continuous-time state equations from discrete-time state equations via the principal  $q$ th root method", IEEE Trans. Automat. Contr., Vol. AC-31, No.5, PP.454-457, 1986
- [3] N.K. Sinha and G.P. Rao, *Identification of continuous-time systems*, Boston: Kluwer Academic, 1991.
- [4] J. Ezzine and C.D. Johnson, "Analysis of continuous/discrete model parameter sensitivity via a perturbation technique", in IEEE 18th Southeastern Symp.on System Theory, PP. 545-550, Knoxville, TN, 1986.
- [5] R.Bellman, *Introduction to Matrix Analysis*, New York: McGraw-Hill, 1982.
- [6] L.S Shieh, Jingfong Gu, and J.W. Sunkel, "Model conversions of continuous and discrete uncertain linear systems using the bilinear and inverse-bilinear approximation method", 1993 American Control Conference, San Francisco, CA, 1993
- [7] L.S. Shieh, H. Wang, and R.E. Yates, "Discrete-continuous model conversion", Appl. Math. Modeling, Vol.4, PP. 449-455, 1980.
- [8] K.S. Narendra and S.S. Tripathi, "Identification and optimization of aircraft dynamics", J. of Aircraft, Vol.10, No.2, PP. 193-199, 1973

- [9] Y.J. Wang, L.S. Shieh, and J.W. Sunkel, "*Robust stabilization, robust performance, and disturbance attenuation for uncertain linear systems*", Computers Math. Applic. Vol.23, No.11, PP. 67-80, 1992.



# Target Tracking and Recognition Using Jump-Diffusion Processes \*

Anuj Srivastava, Robert S. Teichman, Michael I. Miller

Electronic Systems and Signals Research Laboratory  
Department of Electrical Engineering  
Washington University,  
St. Louis, Missouri 63130

## 1 Introduction

In this paper a new approach for target tracking and recognition is presented. We take a Bayesian approach and define a prior density on the scenes of targets and combine it with likelihoods based on the sensor data to give a posterior measure on the parameter space. The jump-diffusion random sampling algorithm [1, 2, 3] is used to sample from the posterior.

In Section 2 the basic approach for solving the problem is described. The global shape model approach is described in section 3 with the Bayesian posterior measure derived in section 4. Section 5 illustrates the use of jump-diffusion processes for random sampling from the posterior measure over the parameter space. The last section describes the implementation on a parallel processing machine and the results obtained.

## 2 Recognition Via Deformable Templates

A fundamental task in the representation of complex dynamically changing scenes involving rigid targets is the construction of models that incorporate both the variability of orientation, range, and object number, as well as the precise rigid structure of the objects in a mathematically precise way. The global deformable shape models introduced in Grenander's general pattern theory [4] extended to parametric representations of arbitrary unknown model order [1, 2] are intended to do this. This becomes the basis for our approach.

There are various kinds of variability and uncertainty inherent to data obtained via remote sensing via high resolution and tracking radars. The first and foremost variability is associated with the conformations of the rigid bodies: orientations, scales, and position. To accommodate this type of variability we use global templates which are made flexible via the introduction of basic transformations involving both rigid motions of translation and rotation, as well as non-rigid motions such as scale. As

these transformations are parameterized by positions in  $\mathbb{R}^3$  and orientations in  $[0, 2\pi)^2 \times [0, \pi]$ , they are performed using continuous stochastic gradient search. The *second* kind of variability is associated with the model order, or parametric dimension. In any scene there may be multiple, variable numbers of targets, with tracks of variable lengths and the target number unknown apriori and, therefore, to be estimated itself.

We take a Bayesian approach by defining a prior density on the scenes of targets. The prior coupled to the sensor data likelihood gives the Bayes posterior. The conformation is selected to be consistent with the data in the sense that scenes of high probability under the posterior distribution are selected. Our method for generating candidate conformations is to sample from the posterior. For this a new class of random sampling algorithms is used based on jump-diffusion dynamics, introduced in Grenander and Miller [1, 2, 3] which visit candidate solutions according to the posterior density. The original motivation for introducing jump-diffusions is to accommodate the very different continuous and discrete components of the discovery process. Given a conformation associated with a target type, or group of targets, the problem is to track and identify the orientation, translation and scale parameters accommodating the variability manifest in the viewing of each object type. For this, the parameter space is sampled using Langevin stochastic differential equations in which the state vector continuously winds through the translation-rotation-scale space following gradients of the posterior. The second distinct part of the sampling process supports the recognition associated with choosing the target types. The deduction algorithm goes through multiple stages of hypothesis during which the airplane types are being discovered, and some subset of the scene may be only partially "recognized." This is accommodated by defining the second transformation type which jumps between different object types, where a jump may correspond to the hypothesis of a new object in the scene, or a "change of mind" about an object type. The jump intensities are governed by the posterior density, with the process visiting conformations of higher probability for longer exponential times, and the diffusion equation governing the dynamics between jumps.

\*This research was supported by ONR N00014-92-J-1418, ARO DAA 03-92-6-0141, Rome Laboratory F30602-92-C-004.

We use the global shape models and pattern theoretic approach introduced by Grenander [5, 4]. As the basic building blocks of the hypotheses define the set of generators  $\mathcal{G}$ , the targets placed at the origin of the *reference coordinates* at a fixed orientation, position, and unit scale.

The fundamental variability in target spaces is accommodated by applying the transformations  $T(\vec{\phi})$ ,  $T(\vec{p})$ ,  $T(s)$  to the templates  $\mathcal{G}$  according to

$$T(\vec{\phi}) : \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_1 & \sin\phi_1 \\ 0 & -\sin\phi_1 & \cos\phi_1 \end{bmatrix} \times \quad (1)$$

$$\begin{bmatrix} \cos\phi_2 & 0 & \sin\phi_2 \\ 0 & 1 & 0 \\ \sin\phi_2 & 0 & \cos\phi_2 \end{bmatrix} \begin{bmatrix} \cos\phi_3 & \sin\phi_3 & 0 \\ -\sin\phi_3 & \cos\phi_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$T(\vec{p}) : \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \begin{bmatrix} x_1 + p_1 \\ x_2 + p_2 \\ x_3 + p_3 \end{bmatrix} \quad (2)$$

$$T(s) : \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad (3)$$

where  $\vec{\phi}$  is the triple of rotation angles associated with the rotation of the viewing sphere about each object,  $\vec{p}$  are the translation parameters in  $\mathbb{R}^3$ , and the  $s$  is the scale parameter in  $\mathbb{R}_+$ . These parameterized transformations operate on the template targets of  $\mathcal{G}$  generating the full target space.

Figure 1 shows one of the 3-D ideal targets used for all of the simulations below. The left panel shows a rendering of the target at the origin, the right panel showing the result of applying one of the transformations.

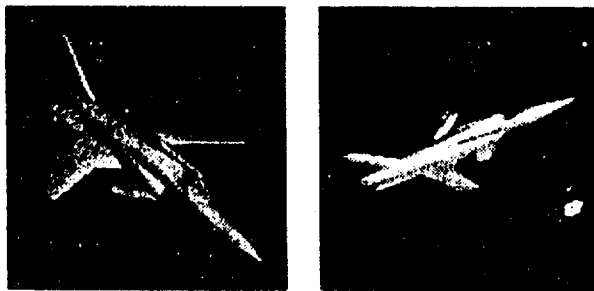


Figure 1: 3-D target at the origin (left panel) and after applying transformations (right panel)

### 3.1 The Parametric Space

Now the parametric space parameterizing the Bayes posterior becomes the set of parameters specifying the similarity transformations, as well as the airplane type. Define the space containing orientations  $\vec{\phi}$  as  $\mathcal{M}(3) \equiv [0, 2\pi)^2 \times [0, \pi]$

with the scale parameter belonging to  $\mathbb{R}_+$ . Then associated with each target or generator  $g \in \mathcal{G}$  is a parameter vector  $\vec{x} \in \mathcal{M}(3) \times \mathbb{R}^3 \times \mathbb{R}_+ \times \mathcal{A}$ , where  $|\mathcal{A}| = |\mathcal{G}|$  the number of different target types.

A pattern will be constructed from multiple targets with varying track lengths. In [6] we have described the multi-target scenario. Here we focus on single target scenes. We are interested in tracking and recognition in "hostile/non-cooperative" environments in which the objects can appear and disappear on random times  $T_1, T + T_1 \in [0, \infty)$  with  $T \geq 0$ . The parameter vector associated with track becomes

$$\vec{x}_T \in \left( \mathcal{M}(3) \times \mathbb{R}^3 \times \mathbb{R}_+ \times \mathcal{A} \right)^{[T_1, T+T_1]} \times \mathbb{R}_+.$$

As tracks will be discretized to sample times  $t_1, t_2, \dots$ , with the object entering and leaving at  $n_1, n + n_1$  respectively, the parameter vector  $\vec{x}_n$  associated with an  $n$ -length track is an element of  $\mathcal{X}_n \equiv \left( \mathcal{M}(3) \times \mathbb{R}^3 \times \mathbb{R}_+ \times \mathcal{A} \right)^n \times \mathcal{Z}_+$ . Since  $n$  is unknown, the full parameter space becomes

$$\mathcal{X} \equiv \bigcup_{n=0}^{\infty} \left( \mathcal{M}(3) \times \mathbb{R}^3 \times \mathbb{R}_+ \times \mathcal{A} \right)^n \times \mathcal{Z}_+. \quad (4)$$

The posterior density defined over the full parameter space  $\mathcal{X}$  is assumed to be of Gibbs form

$$\pi(\vec{x}_n) = \frac{e^{-E(\vec{x}_n)}}{Z}. \quad (5)$$

In the work presented here only rigid transformations are used with  $s = 1$ . Also we assume the entrance time  $T_1(n_1)$  known apriori.

## 4 The Bayes Posterior

The *ideals*  $I \in \mathcal{I}$  are what can be observed by an ideal (with no loss of information) observer. The actual observer, however, may only be able to see the elements with loss of information due to projection, observation noise and/or limited accuracy in the sensor. Denote the operation by which the ideal  $I$  appears as some object say  $I^D$ , by deformations  $d : I \rightarrow I^D$ ,  $d \in \mathcal{D}$ , where  $\mathcal{D}$  is the set of deformation mechanisms, both random and deterministic.

We take a Bayesian approach to the generation of candidate scenes by defining a posterior probability of the parameter vector  $\vec{x}_n$  representing ideal  $I$  given the measured data  $I^D$  according to

$$\pi(\vec{x}_n) = \frac{1}{Z} e^{-E(\vec{x}_n)} \propto \frac{1}{Z} e^{-(P(\vec{x}_n) + L(I^D|\vec{x}_n))},$$

where  $L(I^D|\vec{x}_n)$  is the potential associated with data likelihood and  $P(\vec{x}_n)$  is the potential associated with the prior density on the parameter space.

In the problem stated here the data  $I^D$  has multiple components corresponding to the various sensors:

$$I^D \equiv (I_1^D, I_2^D, \dots).$$

We only include two sensors, one for tracking and one for high-resolution imaging.

#### 4.1 Tracking Priors

The prior on track formation is based on the dynamics of target motion and follows that described in Srivastava et al. [7] in which the force equations governing the motion of targets are utilized to form a prior density on the track parameter space. As an object moves in 3 space it traverses a continuous path consisting of a sequence of translation locations  $\vec{p}(t) \in \mathbb{R}^3$ . For describing their dynamics we shall be interested in expressing the tracks in terms of the body frame velocities of the objects,  $\vec{v}(t) \in \mathbb{R}^3$  which are related to the inertial positions [8] according to

$$\vec{p}(t) = \int_{T_1}^t \Phi(\tau) \vec{v}(\tau) d\tau + \vec{p}(T_1). \quad (6)$$

where  $\Phi(\tau)$  is the product of three rotation matrices in Eqn (2). As in [7] the rigid body analysis with the assumptions that the earth's curvature, motion and wind effects are negligible implies that the translational motion is given by the Newtonian vector equation

$$\dot{\vec{v}}(t) + A(\vec{\phi}(t))\vec{v}(t) = \vec{f}(t). \quad (7)$$

Here  $\vec{\phi}(t) \in \mathcal{M}(3)$  are the Euler's angles representing the orientation of the target with respect to the ground based inertial frame and

$$A(\vec{\phi}(t)) = \begin{bmatrix} 0 & -q_3(t) & q_2(t) \\ q_3(t) & 0 & -q_1(t) \\ -q_2(t) & q_1(t) & 0 \end{bmatrix},$$

with  $\vec{q}(t)$  the rates of change of orientation, which are functions of the Euler's angles.

This linear differential equation is characterized by the time-varying parameter matrix  $A(\vec{\phi}(t))$  and force vector  $\vec{f}(t)$ . The covariance is induced following the approach of Srivastava et al. [7] and Amit et al. [9] by assuming the forcing function is a white process with mean  $\vec{f}(t)$  and a fixed spectral density  $\sigma$ , which then induces a Gaussian process  $\vec{v}(t)$  with mean  $\vec{v}(t)$  and covariance operator determined by the differential operator of Eqn. (7) according to

$$\vec{v}(t) = \int_{T_1}^t e^{-\int_{t_1}^t A(\phi(\tau)) d\tau} \vec{f}(t_1) dt_1 + \vec{v}(T_1) \quad (8)$$

$$K_v(t, s) = \sigma \int_{T_1}^{\min(t, s)} [e^{-\int_{t_1}^t A(\phi(\tau)) d\tau}] [e^{-\int_{t_1}^s A(\phi(\tau)) d\tau}]^\dagger dt_1. \quad (9)$$

Using Eqn (6) the prior density on the airplane positions can be written as

$$K_p(t, s) = \int_{T_1}^t \int_{T_1}^s \Phi(\tau_1) K_v(\tau_1, \tau_2) \Phi^\dagger(\tau_2) d\tau_1 d\tau_2. \quad (10)$$

Notice, this covariance is parameterized by the sequence of airplane orientations  $\vec{\phi}(t), t \in [T_1, T_2]$ . This connects the tracking and recognition algorithms.

## 4.2 The Likelihood: Tracking and Imaging Data

There are two sensor types in our problem: a *tracking sensor* and a *high-resolution imaging sensor*.

### 4.2.1 Tracking

For the tracking we assume a narrowband tracking cross array as in [10, 6, 7, 11] using the standard narrowband signal model developed in [12]. The uniform cross array consists of two uniform linear arrays orthogonal to each other, sensitive to the range, elevation, and azimuth locations of the targets related to the inertial positions  $\vec{p}(t)$  through regular coordinate transformations.

For the data collected at the P-element sensor array at time  $t$  the superposition of the incoming signal and the ambient noise becomes

$$\mathbf{d}_{track}(t) = \mathbf{d}(\vec{p}(t))s(t) + \mathbf{n}(t), \quad (11)$$

where  $\mathbf{n}(t)$  is a  $P \times 1$ , 0-mean complex gaussian random vector with identity covariance,  $s(t)$  is the signal value and  $\mathbf{d}(\vec{p}(t))$  is a regular  $P \times 1$  vandermonde direction vector with the angles of signal arrival parameterized by the inertial position  $\vec{p}(t)$ . The *deterministic signal model* is used as in [10] in which the measurements  $\mathbf{y}(t)$  are Gaussian distributed with mean  $\mathbf{d}(\vec{p}(t))s(t)$ .

### 4.2.2 Imaging

While we are currently incorporating models for high-resolution radar imaging as described in [13, 14, 15, 16, 17, 18], all of the results shown are based on optical imaging systems in which the data is assumed to be on a discrete square  $64 \times 64$  lattice  $\mathcal{L}$ . The imaging data at time  $t$  is the set of 4096 grey scale pixel values  $\mathbf{d}_{recog}(t) \equiv \{d_i(t), i \in \mathcal{L}\}$ . The deformation of the imaging process of the ideal targets is assumed here to be projection with additive noise. For the simulations shown below a Gaussian noise model was used, with the measured data having mean the true ideal 3-D image projected onto the 2-D lattice space.

Shown in Figure 2 are the two kinds of data which the algorithms are based on. The left column shows the ideal projected onto 2-D with additive noise at three different time instants. This is representative of the optical imaging component of the algorithm. The right column shows the spatial power spectrum from the narrowband tracker at three instants of time, plotted in the azimuth-elevation plane (bright is low power, dark is high power).

The two measurements become  $I_1^D \equiv \{\mathbf{d}_{track}(t), t \in [0, \infty)\}$ ,  $I_2^D \equiv \{\mathbf{d}_{recog}(t), t \in [0, \infty)\}$ .

## 5 Jump-Diffusion Random Sampling Algorithm

The most crucial part of the problem still remaining is the derivation of the inference algorithm for generating

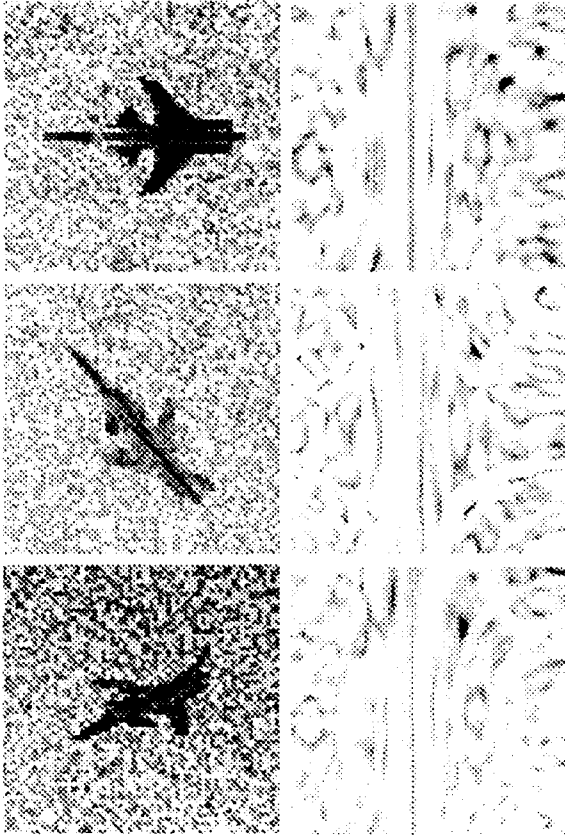


Figure 2: The left column shows the target projected onto the 2-D lattice with additive noise at three different time instants. The right column shows the azimuth-elevation signal power profile at three different instants of time as generated from the narrowband tracking data.

inferences of high probability. Our approach is to construct a jump-diffusion Markov process, following the approach outlined in Grenander and Miller [1], which has the limiting property that it converges in distribution to the Bayes posterior. This jump-diffusion Markov process  $\{X(t), t \geq 0\}$  samples the posterior density  $\pi(\vec{x}_n)$ ,  $\vec{x}_n \in \mathcal{X}$  defined over the full parameter space  $\mathcal{X} = \cup_{n=0}^{\infty} \mathcal{X}_n$ , i.e. the time samples of the process visit the conformations according to the posterior density. This result is presented in [1, 2, 3] as theorems which follow from two fundamental results. First it is shown that  $\pi(\vec{x}_n)$  is a stationary measure of the process  $\{X(t), t \geq 0\}$ , secondly a proof that it is the unique stationary measure is presented. The first part is proven by showing that the backward kolmogoroff operator  $A$  associated with the Markov process satisfies the condition  $\int_{\mathcal{X}} A f(\vec{x}_n) \mu(d\vec{x}_n) = 0$ , for  $f(\cdot) \in \text{domain}(A)$  and  $\mu(d\vec{x}_n) = \pi(\vec{x}_n) m(d\vec{x}_n)$ , the distribution corresponding to  $\pi(\vec{x}_n)$  with  $m(\cdot)$  being the lebesgue measure associated with the parameter space  $\mathcal{X}_n$ . The generator  $A$  has two parts,  $A = A^j + A^d$ , corresponding to the jump and diffusion terms.

## 5.1 Jump Process

The jump process travels through the infinite number of subspaces carrying the inference from subspace to subspace. The two kinds of jump moves allowed here are addition or deletion of track segments from the track configuration. These jump moves are performed based on the intensity parameters derived from relative posterior energies of the configurations. For  $\vec{x}_n$  being the current configuration and  $\vec{y}_m$  a possible candidate,  $\vec{x}_n, \vec{y}_m \in \mathcal{X}$ , define  $q(\vec{x}_n, d\vec{y}_m)$  as the transition intensity,  $q(\vec{x}_n)$  as the intensity of jumping out of the space containing  $\vec{x}_n$ , and  $Q(\vec{x}_n, d\vec{y}_m)$  as the probability of transition. These are related according to the relations

$$q(\vec{x}_n) = \int_{T^1(\vec{x}_n)} q(\vec{x}_n, d\vec{y}_m)$$

and

$$Q(\vec{x}_n, d\vec{y}_m) = \frac{q(\vec{x}_n, d\vec{y}_m)}{q(\vec{x}_n)},$$

where  $T^1(\vec{x}_n)$  is the set of all configurations reachable in one jump move from  $\vec{x}_n$ . As shown in [2] there are at least two ways of generating these jump intensities from the posterior density, these being analogues of Gibbs's sampling and Metropolis [19] based acceptance-rejection. The implementation presented here uses the latter according to which the jump intensity is defined as

$$q(\vec{x}_n, d\vec{y}_m) = e^{-[L(I^D|\vec{y}_m) - L(I^D|\vec{x}_n)]_+} e^{P(\vec{y}_m)} m(d\vec{y}_m) \quad (12)$$

where  $[f(\cdot)]_+$  stands for the positive part of the function. The backward kolmogoroff operator  $A^j$  for this jump process is given by

$$A^j f(\vec{x}_n) = -q(\vec{x}_n) f(\vec{x}_n) + \int_{T^1(\vec{x}_n)} q(\vec{x}_n) Q(\vec{x}_n, d\vec{y}_m) f(\vec{y}_m)$$

and  $\int_{\mathcal{X}} A^j f(\vec{x}_n) \mu(d\vec{x}_n) = 0$ . This makes  $\pi(\vec{x}_n)$  stationary for the jump part of the process.

## 5.2 Diffusion Process

Between jump transitions the diffusion process searches through the uncountable set of parameters within each of the subspaces  $\mathcal{X}_n$ . It is a sample path continuous process which essentially performs a randomized gradient descent over the posterior potential  $E(\vec{x}_n)$  associated with parameter space  $\mathcal{X}_n$  according to Langevin's stochastic differential equation (SDE),

$$dX(t) = \nabla E(X(t)) dt + \sqrt{2} dW(t) \quad (13)$$

where  $W(t)$  is the standard vector wiener process of dimension of the parameter space  $\mathcal{X}_n$ . The backward kolmogoroff operator  $A^d$  defining the diffusion generated by above SDE is given by

$$A^d f(\vec{x}_n) = \nabla E(\vec{x}_n) \cdot \nabla f(\vec{x}_n) + \sum_{i,j} \frac{\partial^2 f(\vec{x}_n)}{\partial(\vec{x}_n)_i \partial(\vec{x}_n)_j}$$

and satisfies the condition  $\int_{\mathcal{X}} A^d f(\vec{x}_n) \mu(d\vec{x}_n) = 0$ . See [1] for the proof.

## 6 Results

The tracking and recognition algorithms were jointly implemented using a Silicon Graphics workstation for data generation and visualization, and a massively parallel 4096 processor SIMD DECmpp machine for implementing the tracking-recognition random sampling algorithm.

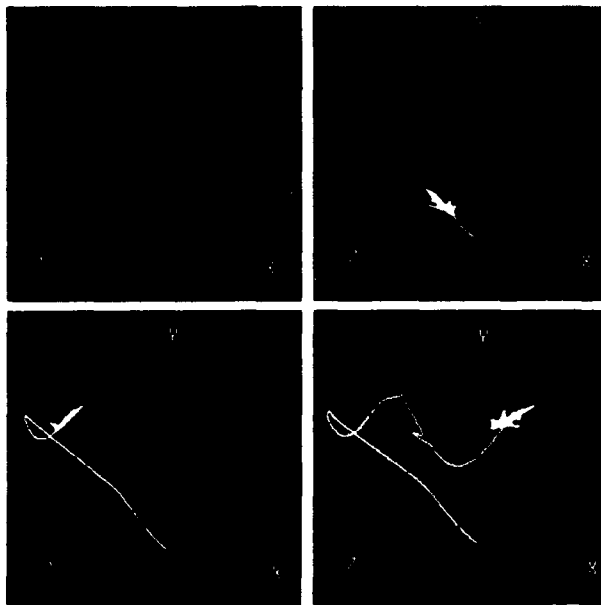


Figure 3: Tracking and recognition using the jump-diffusion algorithm. Top left shows the actual track, the other three panels display the different stages of airplane identification as well as position and orientation estimation.

Using the flight simulator on the Silicon Graphics, a parameterized flight path (Figure 3 top left panel) was generated and stored. From this, tracking data was simulated on the DECmpp assuming a narrowband cross-array of two 32-element uniform linear arrays orthogonal to each other. Optical imaging of the space around the estimated target positions is simulated using the Silicon Graphics to generate data, as shown in Figure 2, which constitutes the imaging data. These two data sets combine to form the posterior density over the parameter space conditioned on the available data up to current time  $t$ . For each candidate target type, a set of "snap-shots" sampling efficiently the object space, all the possible orientations of the object, is pre-stored. The position, orientation and target-type parameters are simultaneously estimated using the jump-diffusion algorithm to sample the posterior density. The estimation proceeds by births and deaths of track segments at random times through discrete jump moves. These moves are performed by generating candidate configurations from the prior density, using target dynamics in this case, and selecting using Metropolis acceptance/rejection. A second type of jump move is allowed to sample from the the object space  $\mathcal{A}$ . A diffusion algorithm is run between

the jumps for adjusting the orientation and position estimates following gradients over the posterior energy. The estimation utilizes the prior measure on target positions which is parameterized by the rotational motion of the target. The object recognition is coupled to the target tracking by use of orientation estimates in the prior measure. In Figure 3 the top left panel shows the actual flight path generated via the flight simulator. The other three panels display the successive stages of the tracking and estimation. The estimated track is superimposed in white on the actual track.

## References

- [1] U. Grenander and M. I. Miller. Jump-diffusion processes for abduction and recognition of biological shapes. *Monograph of the Electronic Signals and Systems Research Laboratory*, 1991.
- [2] U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society*, to appear October, 1993.
- [3] Y. Amit, U. Grenander, and M.I. Miller. Ergodic properties of jump-diffusion processes. *Annals of Applied Probability*, submitted December 1992.
- [4] U. Grenander. Advances in pattern theory: The 1985 rietz lecture. *The Annals of Statistics*, 17:1-30, 1985.
- [5] W. Freiberger and U. Grenander. Computer generated image algebras. *International Federation Information Processing*, 68:1397-1404, 1969.
- [6] A. Srivastava, M. I. Miller, and U. Grenander. Jump-diffusion processes for object tracking and direction finding. In *Proc. 29th Annual Allerton Conference on Communication, Control and Computing*, pages 563-570, Urbana, IL, October 1991. University of Illinois.
- [7] A. Srivastava, N. Cutaia, M. I. Miller, J. A. O'Sullivan, and D. L. Snyder. Multi-target narrow-band direction finding and tracking using motion dynamics. In *Proc. 30th Annual Allerton Conference on Communication, Control, and Computing*, pages 279-288, Urbana, IL, October 1992. University of Illinois.
- [8] Bernard Friedland. *Control System Design : An Introduction To State-Space Methods*. McGraw-Hill Book Company, 1986.
- [9] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *J. American Statistical Association*, 1991.
- [10] M.I. Miller and D. R. Fuhrmann. Maximum likelihood narrow-band direction finding and the em algorithm. *IEEE Acoust. Speech and Signal Processing*, 38, No.9(38, No.9):560-577, 1990.

- [11] M.I. Miller, R. S. Teichman, A. Srivastava, J.A. O'Sullivan, and D. L. Snyder. Jump-diffusion processes for automated tracking-target recognition. In *1993 Conference on Information Sciences and Systems*, Baltimore, Maryland, March 24-26 1993. Johns Hopkins University.
- [12] R. Schmidt. *A signal subspace approach to multiple emitter location and spectral estimation*. Ph.D. Dissertation of Stanford University, Palo Alto, CA., Nov. 1981.
- [13] D.L. Snyder, J.A. O'Sullivan, and M.I. Miller. The use of maximum-likelihood estimation for forming images of diffuse radar-targets. In *Transactions of SPIE in Advanced Architectures and Algorithms*, San Diego, California, 1987.
- [14] D.L. Snyder, J.A. O'Sullivan, and M.I. Miller. The use of maximum-likelihood estimation for forming images of diffuse radar-targets from delay-doppler data. *IEEE Transactions on Information Theory*, pages 536-548, 1989.
- [15] J.A. O'Sullivan, P. Moulin, and D.L. Snyder. Cramer-rao bounds for constrained spectrum estimation with application to a problem in radar imaging. In *Proceedings 26th Allerton Conference on Communication, Control, and Computing*, Champaigne, Urbana, October 1988. Urbana, IL.
- [16] M.I. Miller, D.R. Fuhrmann, J.A. O'Sullivan, and D.L. Snyder. Maximum-likelihood methods for toeplitz covariance estimation and radar imaging. In Simon Haykin, editor, *Advances in Spectrum Estimation*. Prentice-Hall, 1990.
- [17] P. Moulin, J.A. O'Sullivan, and D.L. Snyder. A method of sieves for multiresolution spectrum estimation and radar imaging. *to appear in IEEE Transactions on Information Theory*, 1992.
- [18] J.A. O'Sullivan, K. C. Du, R. S. Teichman, M.I. Miller, D.L. Snyder, and V.C. Vannicola. Radar target recognition using shape models. In *Proc. 30th Annual Allerton Conference on Communication, Control, and Computing*, pages 515-523, Urbana, IL., 1992. University of Illinois.
- [19] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Physical Chemistry*, 21:1087, 1953.

# SUMMARY OF HEAVY TRAFFIC CONVERGENCE OF A CONTROLLED, MULTI-CLASS QUEUEING SYSTEM\*

L. F. Martins, S. E. Shreve, H. M. Soner<sup>†</sup>  
Department of Mathematics  
Carnegie Mellon University

September 1993

## 1 Introduction

In this paper we outline a rigorous proof of the connection between the optimal sequencing problem for a two-station, two-customer-class queueing network and the problem of control of a multi-dimensional diffusion process, obtained as a heavy traffic limit of the queueing problem. Complete proofs are given in a forthcoming paper, Martins, Shreve and Soner (1993). We also describe how to use the diffusion problem, which is one of "singular control" of a Brownian motion (also called "regulated Brownian motion" by Harrison (1985)), to develop policies which are shown to be asymptotically optimal as the traffic intensity approaches one in the queueing network.

The diffusion we wish to control here has been given the name *Brownian network* by Harrison (1988), who proposed such models as approximations to multi-class queueing networks. The control of Brownian networks for the purpose of obtaining control policies for queueing networks was initiated by Wein (1990a, 1990b, 1992) and Harrison & Wein (1989, 1990). These papers derive rules for sequencing cus-

---

\*This work was partially supported by the Army Research Office and the National Science Foundation through the Center for Nonlinear Analysis.

<sup>†</sup>The work of this author was partially supported by the National Science Foundation under grant DMS-9200801.

tomer services and for controlling input to queueing networks. Laws & Louth (1990) and Laws (1991) use Brownian networks to derive queueing network routing policies as well.

All these papers are based on a heuristic understanding, amply supported by simulations, of the connection between the Brownian network control problem and the original queueing problem. Such a connection has been rigorously established only in models with a single customer class, by Kushner & Ramachandran (1988, 1989), Kushner & Martins (1990, 1991), and Krichagina, Lou & Sethi (1993). These papers use weak convergence methods. The exogenous processes (e.g. arrival and service processes) can be quite general, provided they have finite first and second moments.

In this paper, we assume that the arrival processes are Poisson and the service times are exponentially distributed. We base our analysis on the Hamilton-Jacobi-Bellman (HJB) equation, which, in turn, is based on the Markov property. In contrast to other rigorous treatments of convergence, we are able to treat a network with multiple customer classes. Our analysis uses the theory of viscosity solutions of HJB equations. Viscosity solutions were first introduced by Crandall & Lions (1984) and equivalent definitions were given by Crandall, Evans & Lions (1984). For recent developments we refer the reader to Crandall, Ishii & Lions (1992) and Fleming & Soner (1993).

The particular example chosen for our study has been examined by Harrison & Wein (1989) and Chen, Yang & Yao (1991). The former work derives a plausible asymptotically nearly optimal sequencing policy for the queueing network in one of the parameter cases we study; we confirm the asymptotic near-optimality of this policy. The latter work, which does not introduce the Brownian network, solves the original queueing problem in some parameter cases; we obtain consistent results in the case where comparison of results is appropriate, and we obtain an asymptotically nearly optimal policy in a parameter case not solved by Chen, et. al. (1991).

This paper is organized as follows. In Section 2 we describe enough of the queueing system problem, including the heavy traffic assumptions, to enable us to summarize our results. We complete the problem formulation in Section 3. In Section 4 we define the limit of the value functions for a sequence of queueing systems. Of course, our goal is to represent this limit as the value function for a diffusion control problem, and to use this representation to construct asymptotically optimal policies for the queueing systems. In Section 5 we introduce the associated controlled Brownian network, and in Section 6 we reduce the Brownian network problem to one of workload control. Section 7 dispatches the easy case I. Section



8 provides an overview of the harder case II. Full rigorous technical analysis of a subcase of case II, which we call case II A, is given in our forthcoming paper, Martins, Shreve & Soner (1993).

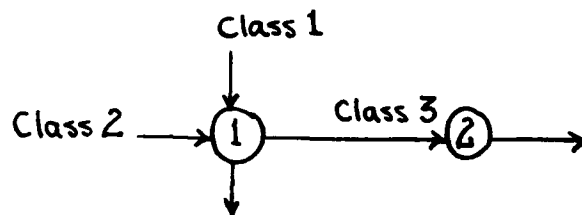
We choose only case II A for full treatment because:

- i) it corresponds to the common situation of seeking to minimize the sum of the queue lengths when the service time at station one is independent of customer class;
- ii) a closed-form solution to the queueing system problem in this subcase is unknown;
- iii) the convergence result in this subcase requires new methodology; and
- iv) the workload control problem in this subcase has a simple solution.

We believe that the techniques developed here can be extended to the other cases and to other problems.

## 2 Summary of results

We study a family of two-station queueing networks with Poisson arrivals and exponential service times. In the  $n^{\text{th}}$  network, customers of class 1 and 2 arrive at station 1 with arrival rates  $\lambda_1^{(n)}$  and  $\lambda_2^{(n)}$ , respectively,



and are served at respective rates  $\mu_1^{(n)}$  and  $\mu_2^{(n)}$ . Class 1 customers then exit the system, whereas class 2 customers proceed to station 2, where they are redesignated as class 3 customers and served at rate  $\mu_3^{(n)}$ .

The cost per unit time of holding a class  $i$  customer is  $c_i > 0$ . The objective is to minimize

$$(2.1) \quad E \int_0^\infty e^{-\alpha t/n} \sum_{i=1}^3 c_i Q_i^{(n)}(t) dt,$$

where  $Q_i^{(n)}$  is the number of class  $i$  customers queued or undergoing service at time  $t$ , and  $\alpha$  is a positive constant.

In order to minimize this objective, we may decide at each time  $t$  whether to serve a class 1 or a class 2 customer. Service can be switched away from one class to the other and subsequently switched back, resuming where it left off. We may also decide to idle station 1, even though there are customers which could be served. This may be desirable if there are no class 1 customers and the cost  $c_3$  is high relative to  $c_2$ , so that we prefer not to serve any class 2 customers until a backlog of class 3 customers has been reduced.

We want these networks to approach heavy traffic conditions as  $n \rightarrow \infty$ . Therefore, we define numbers  $b_1^{(n)}, b_2^{(n)}$  by the formulas

$$(2.2) \quad \frac{\lambda_1^{(n)}}{\mu_1^{(n)}} + \frac{\lambda_2^{(n)}}{\mu_2^{(n)}} = 1 - \frac{b_1^{(n)}}{\sqrt{n}}, \quad \frac{\lambda_2^{(n)}}{\mu_3^{(n)}} = 1 - \frac{b_2^{(n)}}{\sqrt{n}},$$

so that  $1 - b_1^{(n)}/\sqrt{n}$  is the traffic intensity at station 1 and  $1 - b_2^{(n)}/\sqrt{n}$  is the traffic intensity at station 2. The *heavy traffic assumption* is that for  $i = 1, 2, 3$  and  $j = 1, 2$ , the limits

$$\lambda_j = \lim_{n \rightarrow \infty} \lambda_j^{(n)}, \mu_i = \lim_{n \rightarrow \infty} \mu_i^{(n)}, b_j = \lim_{n \rightarrow \infty} b_j^{(n)}$$

are defined, positive, and satisfy

$$(2.3) \quad \sup_n \left[ \sqrt{n} \sum_{j=1}^2 |\lambda_j^{(n)} - \lambda_j| + \sqrt{n} \sum_{i=1}^3 |\mu_i^{(n)} - \mu_i| + \sum_{j=1}^2 |b_j^{(n)} - b_j| \right] < \infty.$$

Our analysis divides naturally into two main cases, and the second case divides into four subcases. We describe our results in each case.

**CASE I:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$ .

As long as customer class 2 is present, it should be served. If all class 2 customers have been served, then class 1 customers should be served.

This result agrees with Theorem (5.2) of Chen, Yang & Yao (1991). The expected cost reduction per unit of service effort devoted to a class 2 customer is  $(c_2 - c_3)\mu_2$ , since service turns a class 2 customer into a class 3 customer. In Case I,  $(c_2 - c_3)\mu_2$  dominates  $c_1\mu_1$ , the expected cost reduction per unit of service effort to a class 1 customer. This results in the simple fixed priority rule of serving class 2 customers whenever they are present.

**CASE II:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ .

We further divide Case II into four subcases.

**CASE II A:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ ,  $c_2\mu_2 - c_3\mu_2 \geq 0$ ,  $c_2\mu_2 - c_1\mu_1 \geq 0$ .

Now a unit of service applied to class 1 results in a greater expected cost reduction than a unit of service to class 2. We prove the near asymptotic optimality of the policy of serving class 1 unless the number of class 3 customers falls below a positive threshold, in which case priority is switched to class 2, so that station 2 is not starved. (The notion of near asymptotic optimality is defined in (4.4) below.) The nearly optimal policy depends on all the customers present, and it is given by,

$$\begin{aligned} \text{serve class 1 if } & \gamma\left(\frac{1}{\sqrt{n}} Q_1^{(n)}(t), \frac{1}{\sqrt{n}} Q_2^{(n)}(t), \frac{1}{\sqrt{n}} Q_3^{(n)}(t)\right) \geq 0, \\ \text{serve class 2 if } & \gamma\left(\frac{1}{\sqrt{n}} Q_1^{(n)}(t), \frac{1}{\sqrt{n}} Q_2^{(n)}(t), \frac{1}{\sqrt{n}} Q_3^{(n)}(t)\right) < 0, \end{aligned}$$

where  $Q_i^{(n)}(t)$ ,  $i = 1, 2, 3$ , denotes the number of class- $i$  customers present at time  $t$ ,  $\gamma$  is any function of the form

$$\gamma(z_1, z_2, z_3) = a(z_1)a(z_3) - b(z_2),$$

and  $a$  and  $b$  are any bounded, nonnegative, concave, increasing functions satisfying,

$$a(0) = b(0) = 0, \quad b(\infty) < a(\infty)^2.$$

In Martins, Shreve & Soner (1993), we show that as  $b(\infty)$  goes to zero, the above policy becomes asymptotically optimal. Harrison & Wein's (1989) model with  $c_1 = c_2 = c_3 = 1$ ,  $\mu_1 = \mu_2 = 2$ ,  $\mu_3 = 1$  falls into this subcase, and their proposed policy is to serve class 1 if and only if  $\sqrt{n} Q_2^{(n)}(t)$  exceeds a positive constant which is independent of  $n$  and the other queue lengths. They showed by simulation that with a properly chosen constant, this policy outperforms the rules "first-in first-out," "longest expected remaining processing time," and "shortest expected remaining

processing time." They also found that its performance was close (within about 5%) of a lower bound they obtained for optimal performance.

The heuristic justification of the policy is case II A suggests that the same policy is asymptotically optimal under only the case II condition  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ . Our proof of the result stated in case II A suggests otherwise. Although we have not worked out a full proof for the other three subcases, the proof for case II A strongly suggests the following conjectures.

**CASE II B:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0, c_2, \mu_2 - c_3\mu_2 < 0, c_2\mu_2 - c_1\mu_1 \geq 0$ .

There is a continuous, increasing function  $\Psi_2 : [0, \infty) \rightarrow [0, \infty)$  satisfying,

$$0 \leq \Psi_2(\omega_2) < \mu_3\omega_2/\omega_1, \quad \forall \omega_2 \geq 0, \quad \text{and} \quad \lim_{\omega_2 \rightarrow \infty} \Psi_2(\omega_2) = \infty.$$

Class 1 should be given priority, unless either the queue length  $Q_1^{(n)}$  of class 1 customers falls to zero or the queue length  $Q_3^{(n)}$  of class 3 customers falls below some small threshold. While either of these conditions is satisfied, priority should be switched to class 2, except that whenever  $Q_1^{(n)} = 0$  and  $Q_2^{(n)} < \sqrt{n} \mu_2 \Psi_1((Q_2^{(n)} + Q_3^{(n)}/\sqrt{n} \mu_3))$ , station 1 should be idled. This idleness can be explained by the fact that it is cheaper to hold class 2 customers at station 1 than to send them on to be held as class 3 customers at station 2; note that in this subcase,  $c_2 < c_3$ .

**CASE II C:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0, c_2\mu_2 - c_3\mu_2 \geq 0, c_2\mu_2 - c_1\mu_1 < 0$ .

There exists a continuous, increasing function  $\Psi_1 : [0, \infty) \rightarrow [0, \infty)$  satisfying,

$$0 \leq \Psi_1(\omega_1) < \mu_2\omega_1/\omega_2, \quad \forall \omega_1 \geq 0, \quad \text{and} \quad \lim_{\omega_1 \rightarrow \infty} \Psi_1(\omega_1) = \infty.$$

Class 1 should be given priority unless either  $Q_1^{(n)} = 0$  or  $Q_3^{(n)}$  is less than a small threshold. While either of these conditions is satisfied, priority should be switched to class 2, except that when  $Q_1^{(n)} > 0$ , and

$$Q_2^{(n)} < \sqrt{n} \mu_3 \Psi_1((Q_1^{(n)}/\sqrt{n} \mu_1) + (Q_2^{(n)}/\sqrt{n} \mu_2)),$$

priority should be given to class 1, even though this may cause station 2 to starve. Idling station 2 can be explained by the fact that the cost of operating the network

can be reduced more quickly by serving class 1 than by serving class 2; note that  $c_1\mu_1 > c_2\mu_2$ .

**CASE II D:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ ,  $c_2\mu_2 - c_3\mu_2 < 0$ ,  $c_2\mu_2 - c_1\mu_1 < 0$ .

This case is a combination of case II B and case II C. We conjecture the existence of functions  $\Psi_1$  and  $\Psi_2$  as described above. Idling can occur at either station 1 or station 2, as described in case II B and case II C, respectively.

### 3 Queueing network problem

For the queueing network of the previous section, for  $i = 1, 2$ , let  $\{A_i^{(n)}(t); 0 \leq t < \infty\}$  be the class  $i$  customer *arrival process*, assumed to be Poisson with intensity  $\lambda_i^{(n)}$ . For  $i = 1, 2, 3$ , let  $\{S_i^{(n)}(t); 0 \leq t < \infty\}$  be the class  $i$  customer *service process*, assumed to be Poisson with intensity  $\mu_i^{(n)}$ . We take all these processes to be left-continuous, and we denote by  $\{\mathcal{F}^{(n)}(t); 0 \leq t < \infty\}$  the filtration generated by these five processes.

A *control law*  $\{Y(t), U(t); 0 \leq t < \infty\}$  is a pair of left-continuous,  $\{\mathcal{F}^{(n)}(t); 0 \leq t < \infty\}$ -adapted,  $\{0, 1\}$ -valued processes. The process  $Y(t)$  indicates whether station 1 is active ( $Y(t) = 1$ ), or idle ( $Y(t) = 0$ ), and  $U(t)$  indicates whether station 1 is serving customer class 1 ( $U(t) = 1$ ) or customer class 2 ( $U(t) = 0$ ). Given non-negative initial queue lengths  $Q_1^{(n)}(0)$ ,  $Q_2^{(n)}(0)$  and  $Q_3^{(n)}(0)$  for the three customer classes, and given a control law  $(Y, U)$ , we define the *queue length processes*

$$Q_1^{(n)}(t) = Q_1^{(n)}(0) + A_1^{(n)}(t) - \int_0^t Y(s)U(s)1_{\{Q_1^{(n)}(s) \geq 1\}} dS_1^{(n)}(s),$$

$$Q_2^{(n)}(t) = Q_2^{(n)}(0) + A_2^{(n)}(t) - \int_0^t Y(s)(1 - U(s))1_{\{Q_2^{(n)}(s) \geq 1\}} dS_2^{(n)}(s),$$

$$Q_3^{(n)}(t) = Q_3^{(n)}(0) + \int_0^t Y(s)(1 - U(s))1_{\{Q_2^{(n)}(s) \geq 1\}} dS_2^{(n)}(s), \\ - \int_0^t 1_{\{Q_3^{(n)}(s) \geq 1\}} dS_3^{(n)}(s).$$

We denote the vector of queue length processes by

$$Q^{(n)}(t) = (Q_1^{(n)}(t), Q_2^{(n)}(t), Q_3^{(n)}(t)).$$

(Note: Because the inter-service times are exponentially distributed, the processes

$$\int_0^t Y(s)U(s)1_{\{Q_1^{(n)}(s) \geq 1\}} dS_1^{(n)}(s) \text{ and } S_1^{(n)}(\int_0^t Y(s)U(s)1_{\{Q_1^{(n)}(s) \geq 1\}} ds)$$

have the same law. This permits us to write  $Q_1^{(n)}(t)$  in terms of the former, although the latter more nearly reflects the way we interpret the system. If service of a customer is preempted and later resumed, we assume that service begins where it was left off. After resumption of service, the time to completion has the same exponential distribution as the original distribution of the service time. Similar comments apply to  $Q_2^{(n)}(t)$  and  $Q_3^{(n)}(t)$ .)

The vector of *scaled queue length processes* is

$$Z^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} Q^{(n)}(nt).$$

This is a Markov chain with lattice state space  $L^{(n)} \triangleq \{\frac{k}{\sqrt{n}}; k = 0, 1, \dots\}^3$ , and, for fixed  $(y, u) \in \{0, 1\}^2$ , its infinitesimal generator is

$$\begin{aligned} (\mathcal{L}^{n,y,u}\varphi)(z) \triangleq & n\lambda_1^{(n)}[\varphi(z + \frac{1}{\sqrt{n}}e_1) - \varphi(z)] + n\lambda_2^{(n)}[\varphi(z + \frac{1}{\sqrt{n}}e_2) - \varphi(z)] \\ & + n\mu_1^{(n)}yu[\varphi(z - \frac{1}{\sqrt{n}}e_1) - \varphi(z)]1_{\{z_1 > 0\}} \\ & + n\mu_2^{(n)}y(1-u)[\varphi(z - \frac{1}{\sqrt{n}}e_2 + \frac{1}{\sqrt{n}}e_3) - \varphi(z)]1_{\{z_2 > 0\}} \\ & + n\mu_3^{(n)}[\varphi(z - \frac{1}{\sqrt{n}}e_3) - \varphi(z)]1_{\{z_3 > 0\}}, \end{aligned} \quad (3.1)$$

where  $z = (z_1, z_2, z_3)$ ,  $e_1 = (1, 0, 0)$ ,  $e_2 = (0, 1, 0)$  and  $e_3 = (0, 0, 1)$ . In particular, given any control law  $(Y(\cdot), U(\cdot))$ , for any real-valued function  $\varphi$  on  $L^{(n)}$ , the process

$$(3.2) \quad e^{-\alpha t} \varphi(Z^{(n)}(t)) + \int_0^t e^{-\alpha s} [\alpha \varphi(Z^{(n)}(s)) - \mathcal{L}^{n,Y(s),U(s)} \varphi(Z^{(n)}(s))] ds$$

is a local martingale.

Using the positive holding costs  $c_1, c_2, c_3$ , we define the *holding cost function*  $h(z) = \sum_{i=1}^3 c_i z_i$ . Given an initial condition  $Z^{(n)}(0) = z \in L^{(n)}$  and a control law  $(Y(\cdot), U(\cdot))$ , we define the associated *cost function* at  $z$  by

$$(3.3) \quad J_{Y,U}^{(n)}(z) \triangleq E \int_0^\infty e^{-\alpha t} h(Z^{(n)}(t)) dt.$$

In terms of the original queue length process, this cost can be written as (cf.(2.1))

$$n^{-\frac{3}{2}} E \int_0^\infty e^{-(\alpha t)/n} h(Q^{(n)}(t)) dt.$$

The value function at  $z$  is

$$(3.4) \quad J_*^{(n)}(z) \triangleq \inf \{ J_{Y,U}^{(n)}(z); (Y, U) \text{ is a control law} \}.$$

For  $\varphi : L^{(n)} \rightarrow \mathcal{R}$ , we define the nonlinear operator  $\mathcal{L}^{n,*}$  acting on  $\varphi$  by

$$(3.5) \quad \mathcal{L}^{n,*} \varphi(z) \triangleq \min \{ \mathcal{L}^{n,y,u} \varphi(z); (y, u) \in \{0, 1\}^2 \} \quad \forall z \in L^{(n)}.$$

The *Hamilton-Jacobi-Bellman* (HJB) equation for the  $n^{\text{th}}$  queueing network is

$$(3.6) \quad \alpha \varphi - \mathcal{L}^{n,*} \varphi - h = 0 \quad \text{on } L^{(n)}.$$

The following is a standard verification theorem.

**Proposition 3.1** *The value function  $J_*^{(n)}$  is the unique, linearly growing solution of the HJB equation (3.6). If  $\varphi$  is a linearly growing subsolution (respectively, supersolution) of this equation, then  $\varphi \leq J_*^{(n)}$  (respectively,  $\varphi \geq J_*^{(n)}$ ). Furthermore, any stationary control law  $(Y^*, U^*)$  satisfying*

$$(3.7) \quad \mathcal{L}^{n,Y^*,U^*} J_*^{(n)} = \mathcal{L}^{n,*} J_*^{(n)}$$

*is optimal.*

## 4 The heavy traffic limit of the value function

In order to let  $n \rightarrow \infty$ , we need an upper bound, independent of  $n$ , for the non-negative functions  $\{J_*^{(n)}\}_{n=1}^\infty$ . The following estimate is an easy consequence of the maximum principle, see Martins, Shreve & Soner (1993), Theorem 6.1.

**Proposition 4.1** *There are constants  $K_1$  and  $K_2$ , independent of  $n$ , such that*

$$J_*^{(n)}(z) \leq K_1 + K_2(z_1 + z_2 + z_3) \quad \forall z \in L^{(n)}.$$

We wish to consider  $\lim_{n \rightarrow \infty} J_*^{(n)}$ , but since each  $J_*^{(n)}$  is defined on a different set  $L^{(n)}$ , the definition of this limit is not straightforward. Borrowing the technique developed by Barles & Perthame (1988) (also see Fleming & Soner (1993, Section 7.3)), we define the *upper semicontinuous limit*  $J^\#$  of  $\{J_*^{(n)}\}_{n=1}^\infty$  by

$$(4.1) \quad J^\#(z) \triangleq \lim_{\epsilon \downarrow 0} \lim_{n \rightarrow \infty} \sup \{J_*^{(n)}(\zeta); \|\zeta - z\| < \epsilon, \zeta \in L^{(n)}\} \quad \forall z \in [0, \infty)^3,$$

and the *lower semicontinuous limit*  $J_\#$  by

$$(4.2) \quad J_\#(z) \triangleq \lim_{\epsilon \downarrow 0} \lim_{n \rightarrow \infty} \inf \{J_*^{(n)}(\zeta); \|\zeta - z\| < \epsilon, \zeta \in L^{(n)}\} \quad \forall z \in [0, \infty)^3.$$

Then  $J^\#$  is upper semicontinuous,  $J_\#$  is lower semicontinuous, and

$$(4.3) \quad 0 \leq J_\#(z) \leq J^\#(z) \leq K_1 + K_2(z_1 + z_2 + z_3) \quad \forall z \in [0, \infty)^3.$$

In (Martins, Shreve & Soner (1993)), we prove that  $J_\# = J^\#$ , and also we use a Brownian network problem to suggest, for each  $\eta > 0$ , a sequence of stationary policies  $\{(Y^{(n)}, U^{(n)})\}_{n=1}^\infty$  such that

$$(4.4) \quad \lim_{\epsilon \downarrow 0} \lim_{n \rightarrow \infty} \sup \{ |J_{Y_n, U_n}^{(n)}(\zeta) - J_\#(\zeta)|; \|\zeta - z\| < \epsilon, \zeta \in L^{(n)} \} \leq \eta,$$

for all  $z \in [0, \infty)^3$ . We call such a family (parametrized by  $\eta$ ) of sequences of policies *asymptotically nearly optimal*.

## 5 The controlled Brownian network.

We first introduce the controlled Brownian network and then explain by an analysis of the infinitesimal generator  $\mathcal{L}^{n,v,u}$  why it is relevant. Let  $M_1, M_2$  and  $M_3$  be continuous martingales relative to a filtration  $\{\mathcal{F}(t)\}$  satisfying the usual conditions



that each  $\mathcal{F}(t)$  contains all null sets of  $\mathcal{F}(\infty)$  and that  $\mathcal{F}(t) = \cap_{s>t} \mathcal{F}(s)$  for all  $t$ . Assume that for all  $t$ ,

$$(5.1) \quad \langle M_1 \rangle(t) = 2\lambda_1 t, \quad \langle M_2 \rangle(t) = \langle M_3 \rangle(t) = 2\lambda_2 t,$$

$$(5.2) \quad \langle M_1, M_2 \rangle(t) = \langle M_1, M_3 \rangle(t) = 0, \quad \langle M_2, M_3 \rangle(t) = -\lambda_2 t.$$

Given  $z \in [0, \infty)^3$ , we will say that the quadruple  $(\ell_0, \ell_1, \ell_2, \ell_3)$  of  $\{\mathcal{F}(t)\}$ -adapted processes is *admissible for initial condition  $z$*  provided:

- (i)  $(\ell_0, \ell_1, \ell_2, \ell_3)$  are right-continuous with left-hand limits, with the convention that  $\ell_i(0^-) = 0$ ,  $i = 1, 2, 3$ ;
- (ii)  $\ell_0$  is of finite variation on bounded intervals;
- (iii)  $\ell_1, \ell_2, \ell_3$  are nondecreasing,
- (iv) the state process  $Z(t) = (Z_1(t), Z_2(t), Z_3(t))$  is in  $[0, \infty)^3$  for all  $t \geq 0$ ,

where

$$(5.3) \quad Z_1(t) \triangleq z_1 + M_1(t) + \mu_1 \ell_0(t) + \ell_1(t),$$

$$(5.4) \quad Z_2(t) \triangleq z_2 - b_1 \mu_2 t + M_2(t) - \mu_2 \ell_0(t) + \ell_2(t),$$

$$(5.5) \quad Z_3(t) \triangleq z_3 + (b_1 \mu_2 - b_2 \mu_3)t + M_3(t) + \mu_2 \ell_0(t) - \ell_2(t) + \ell_3(t).$$

The *cost function* associated with  $(\ell_0, \ell_1, \ell_2, \ell_3)$ , admissible at  $z \in [0, \infty)^3$ , is

$$V_{\ell_0, \ell_1, \ell_2, \ell_3}(z) \triangleq E \int_0^\infty e^{-\alpha t} h(Z(t)) dt.$$

The *value function* for the controlled Brownian network is

$$(5.6) \quad V(z) \triangleq \inf \{V_{\ell_0, \ell_1, \ell_2, \ell_3}(z); (\ell_0, \ell_1, \ell_2, \ell_3) \text{ is admissible at } z\}, \quad z \in [0, \infty)^3.$$

The cross variation formulas (5.1), (5.2) imply that the vector of martingales  $(M_1, M_2, M_3)$  is nothing more than a three-dimensional standard Brownian motion multiplied by a non-singular matrix, so this vector of martingales is also a Markov process. If we set the control processes  $\ell_0, \ell_1, \ell_2, \ell_3$  equal to zero, the state process  $Z(t)$  given by (5.3)-(5.5) is Markov with infinitesimal generator

$$(5.7) \quad \mathcal{L}\varphi = -b_1 \mu_2 \varphi_2 + (b_1 \mu_2 - b_2 \mu_3) \varphi_3 + \lambda_1 \varphi_{11} + \lambda_2 \varphi_{22} - \lambda_2 \varphi_{23} + \lambda_2 \varphi_{33},$$

where  $\varphi$  is any real valued,  $C^2$  function on  $[0, \infty)^3$  with  $\varphi_i$  denoting partial derivative with respect to the  $i^{\text{th}}$  variable.

The controlled Brownian network is an intermediate problem between the queueing networks studied thus far and the workload control problem of the next section. Although the value function is well-defined by (5.6), the problem does not have an optimal solution. One would like to keep the state  $Z(t)$  on a face of the orthant  $[0, \infty)^3$ , but this is not possible with the bounded variation control processes  $\ell_0, \ell_1, \ell_2, \ell_3$ . Fortunately, when we pass to the workload formulation, we will obtain a well-posed control problem.

We conclude this section with an asymptotic expansion of the infinitesimal generator  $\mathcal{L}^{n,y,u}$  of (3.1) for the controlled queueing network. This expansion explains the origin of the Brownian network problem introduced in this section.

Suppose  $\varphi : [0, \infty)^3$  is thrice continuously differentiable, and all derivatives of  $\varphi$  up to order three are bounded uniformly on  $[0, \infty)^3$ . Fix  $(y, u) \in \{0, 1\}^2$ , and define

$$(5.8) \quad \theta = \sqrt{n} \left[ \frac{\lambda_1^{(n)}}{\mu_1^{(n)}} - u \right], \sigma_1 = \sqrt{n} \mu^{(n)} (1 - y) u, \sigma_2 = \sqrt{n} \mu_2^{(n)} (1 - y) (1 - u).$$

Recalling (2.2), we may write

$$(5.9) \quad u = -\frac{\theta}{\sqrt{n}} + \frac{\lambda_1^{(n)}}{\mu_1^{(n)}}, \quad (1 - u) = \frac{\theta}{\sqrt{n}} + \frac{\lambda_2^{(n)}}{\mu_2^{(n)}} + \frac{b_1^{(n)}}{\sqrt{n}}, \quad \lambda_2^{(n)} = \mu_3^{(n)} - \frac{b_2^{(n)} \mu_3^{(n)}}{\sqrt{n}}.$$

For  $z \in [0, \infty)^3$ , we set

$$(5.10) \quad B_1^{n,y,u} \varphi(z) \triangleq \mu_1^{(n)} y u [\sqrt{n} \varphi_1(z) - \frac{1}{2} \varphi_{11}(z)] 1_{\{z_1=0\}},$$

$$(5.11) \quad B_2^{n,y,u} \varphi(z) \triangleq \mu_2^{(n)} y (1 - u) \sqrt{n} B_2^n \varphi(z) 1_{\{z_2=0\}},$$

$$B_2^n \varphi \triangleq \varphi_2 - \varphi_3 - \frac{1}{2\sqrt{n}} (\varphi_{22} + \varphi_{33} - 2\varphi_{23}),$$

$$(5.12) \quad B_3^{n,y,u} \varphi \triangleq \mu_3^{(n)} [\sqrt{n} \varphi_3(z) - \frac{1}{2} \varphi_{33}(z)] 1_{\{z_3=0\}},$$

so that (3.1) becomes (see Martins, Shreve & Soner (1993) for details),

$$\begin{aligned}
\mathcal{L}^{n,y,u}\varphi(z) &= \mathcal{L}\varphi(z) + \theta[\nabla\varphi(z) \cdot \xi^{(n)} + \frac{1}{\sqrt{n}}\mathcal{A}^n\varphi(z)] \\
(5.13) \quad &+ \sigma_1[\varphi_1(z) - \frac{1}{2\sqrt{n}}\varphi_{11}(z)] + \sigma_2 B_2^n \varphi(z) \\
&+ \sum_{i=1}^3 B_i^{n,y,u}\varphi(z) + O(\frac{1}{\sqrt{n}}),
\end{aligned}$$

where  $\mathcal{L}\varphi$  is given by (5.7) and

$$(5.14) \quad \xi^{(n)} \triangleq (\mu_1^{(n)}, -\mu_2^{(n)}, \mu_3^{(n)})$$

$$(5.15) \quad \mathcal{A}^n\varphi \triangleq -\frac{1}{2}\mu_1^{(n)}\varphi_{11} + \frac{1}{2}\mu_2^{(n)}(\varphi_{22} - 2\varphi_{23} + \varphi_{33}).$$

The expressions in (5.14), (5.15) are bounded uniformly in  $n$ . However,  $\theta$ ,  $\sigma_1$  and  $\sigma_2$  are of order  $\sqrt{n}$ , as are the terms  $B_i^{n,y,u}\varphi$ . The term  $\nabla\varphi \cdot \xi^{(n)} + \frac{1}{\sqrt{n}}\mathcal{A}^n\varphi$  in (5.13) agrees with  $\nabla\varphi \cdot \xi$  up to an error of order  $\frac{1}{\sqrt{n}}$ , but this term cannot immediately be replaced by  $\nabla\varphi \cdot \xi$  because  $\theta$  multiplying it is of order  $\sqrt{n}$ .

Equation (5.13) suggests that the controlled Brownian motion  $Z(t)$  given by (5.3)-(5.5) approximates the scaled queue length process  $Z^{(n)}(t) = \frac{1}{\sqrt{n}}Q^{(n)}(nt)$ . The control variable  $\theta$  in (5.13), which can be either positive or negative, corresponds to pushing in approximately the direction  $\xi \triangleq (\mu_1, -\mu_2, \mu_2)$  or the direction  $-\xi$ . In (5.3)-(5.5), this pushing is accomplished by the locally finite-variation process  $\ell_0$ . The processes  $\ell_1, \ell_2$  and  $\ell_3$  appearing in (5.3)-(5.5) allow us to enforce the condition  $Z(t) \in [0, \infty)^3$  for all  $t \geq 0$ . We have set up the controlled Brownian network to allow  $\ell_i$  to grow even when  $Z_i(t) > 0$ ; this corresponds to idling the serving stations.

## 6 The workload formulation

Following Harrison & Wein (1989), we introduce the *workload transformation*

$$(6.1) \quad w_1 = \frac{z_1}{\mu_1} + \frac{z_2}{\mu_2}, \quad w_2 = \frac{z_2 + z_3}{\mu_3},$$

which maps the state space  $[0, \infty)^3$  of the controlled Brownian network onto the state space  $[0, \infty)^2$  of the *workload control problem* formulated in this section. If  $(z_1, z_2, z_3)$  represents the three queue lengths, then  $(w_1, w_2)$  is the expected impending service time for the two stations embodied in customers anywhere in the network. The

workload formulation reduces the dimensionality of the control problem from three (the number of customer classes) to two (the number of stations).

Because we can use the control process  $\ell_0$  in (5.3)-(5.5) to instantaneously change the state  $Z(t)$  in the directions  $\pm \xi \triangleq \pm(\mu_1, -\mu_2, \mu_2)$  at no cost, the Brownian network value function  $V$  of (5.6) will be constant along the direction  $\xi$ . This means that  $V(z_1, z_2, z_3)$  can be written as a function of  $(w_1, w_2)$ , because  $(w_1, w_2)$  does not change along the  $\xi$ -direction. It also means that one would want to keep the process  $Z(t)$  on the locus of points in  $[0, \infty)^3$  which minimize  $h$  along line segments parallel to  $\xi$ . To find this locus, one considers for each  $(w_1, w_2) \in [0, \infty)^2$  the linear program

$$\begin{aligned} &\text{Minimize} && c_1 z_1 + c_2 z_2 + c_3 z_3 \\ &\text{Subject to} && \frac{z_1}{\mu_1} + \frac{z_2}{\mu_2} = w_1, \\ & && \frac{z_2}{\mu_3} + \frac{z_3}{\mu_3} = w_2, \\ & && z_1 \geq 0, \quad z_2 \geq 0, \quad z_3 \geq 0. \end{aligned}$$

Denote by  $\hat{h}(w_1, w_2)$  the value of this linear program. We have two major cases:

**CASE I:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$ .

In this case,

$$(6.2) \quad \hat{h}(w_1, w_2) = c_1\mu_1 w_1 + c_3\mu_3 w_2,$$

and the minimizer in the linear program is

$$(6.3) \quad z_1^* = \mu_1 w_1, \quad z_2^* = 0, \quad z_3^* = \mu_3 w_2.$$

**CASE II:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ .

Now

$$(6.4) \quad \hat{h}(w_1, w_2) = \begin{cases} (c_2\mu_2 - c_3\mu_2)w_1 + c_3\mu_3 w_2 & \text{if } \mu_3 w_2 \geq \mu_2 w_1, \\ c_1\mu_1 w_1 + \frac{\mu_3}{\mu_2}(c_2\mu_2 - c_1\mu_1)w_2 & \text{if } \mu_3 w_2 \leq \mu_2 w_1. \end{cases}$$

The minimizing values are

$$(6.5) \quad \begin{aligned} z_1^* &= 0, \quad z_2^* = \mu_2 w_1, \quad z_3^* = \mu_3 w_2 - \mu_2 w_1 \quad \text{if } \mu_3 w_2 \geq \mu_2 w_1, \\ z_1^* &= \frac{\mu_1}{\mu_2}(\mu_2 w_1 - \mu_3 w_2), \quad z_2^* = \mu_3 w_2, \quad z_3^* = 0 \quad \text{if } \mu_3 w_2 \leq \mu_2 w_1. \end{aligned}$$

The workload control problem has state equations

$$(6.6) \quad W_1(t) = w_1 - b_1 t + \frac{1}{\mu_1} M_1(t) + \frac{1}{\mu_2} M_2(t) + m_1(t),$$

$$(6.7) \quad W_2(t) = w_2 - b_2 t + \frac{1}{\mu_3} M_2(t) + \frac{1}{\mu_3} M_3(t) + m_2(t),$$

where the pair  $(m_1, m_2)$  of  $\{\mathcal{F}(t)\}$ -adapted control processes is admissible for initial condition  $w = (w_1, w_2) \in [0, \infty)^2$  provided:

(i)  $m_1, m_2$  are right-continuous with left-hand limits, with the convention that

$$m_i(0^-) = 0, \quad i = 1, 2;$$

(ii)  $m_1, m_2$  are nondecreasing;

(iii) the state process  $W(t) = (W_1(t), W_2(t))$  is in  $[0, \infty)^2$  for all  $t \geq 0$ .

(We have in mind, of course, that  $m_1(t) = \frac{\ell_1(t)}{\mu_1}$ ,  $m_2(t) = \frac{\ell_2(t)}{\mu_3}$ , where  $\ell_1, \ell_3$  are part of an admissible quadruple  $(\ell_0, \ell_1, \ell_2, \ell_3)$  for the controlled Brownian network.) The cost function associated with  $(m_1, m_2)$  at  $w \in [0, \infty)^2$  is

$$\hat{V}_{m_1, m_2}(w) \triangleq E \int_0^\infty e^{-\alpha t} h(W(t)) dt,$$

and the value function at  $w$  is

$$(6.8) \quad \hat{V}(w) = \inf \{ \hat{V}_{m_1, m_2}(z); (m_1, m_2) \text{ is admissible at } w \}.$$

Although we do not need this fact for our analysis, one can show that  $V$  of (5.6) and  $\hat{V}$  of (6.8) are related by the equation

$$(6.9) \quad V(z_1, z_2, z_3) = \hat{V}\left(\frac{z_1}{\mu_1} + \frac{z_2}{\mu_2}, \frac{z_2 + z_3}{\mu_3}\right) \quad \forall z \in [0, \infty)^3.$$

If one had an optimal  $(m_1^*, m_2^*)$  for the workload control problem, then as an optimal policy for the Brownian network problem, one would want to take  $\ell_1^*(t) = \mu_1 m_1^*(t)$ ,  $\ell_2^*(t) \equiv 0$ ,  $\ell_3^*(t) = \mu_3 m_2^*(t)$ , and choose  $\ell_0$  to ensure that  $Z^*(t)$  is always given by (6.3) or (6.5) with  $w_i = W_i^*(t)$ ,  $i = 1, 2$ , depending on the sign of  $c_1 \mu_1 - c_2 \mu_2 + c_3 \mu_2$ . However, such an  $\ell_0$  does not exist, and so the Brownian network control problem is ill-posed.

## 7 Solution of Case I:

This is the case  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$ . Since  $\hat{h}$  given by (6.2) is increasing in each variable separately, the optimal control processes  $m_1$  and  $m_2$  act only when  $W_1 = 0$  or  $W_2 = 0$ , respectively. More precisely,

$$(7.1) \quad m_1(t) \triangleq \max_{0 \leq s \leq t} [-w_1 + b_1 s - \frac{1}{\mu_1} M_1(s) - \frac{1}{\mu_2} M_2(s)]^+,$$

$$(7.2) \quad m_2(t) \triangleq \max_{0 \leq s \leq t} [-w_2 + b_2 s - \frac{1}{\mu_3} M_2(s) - \frac{1}{\mu_3} M_3(s)]^+$$

are the minimal nondecreasing processes which ensure that the associated state processes remain nonnegative almost surely. In particular,

$$(7.3) \quad m_i(t) = \int_0^t 1_{\{W_i(s)=0\}} dm_i(s), 0 \leq t < \infty.$$

One can actually compute the value function

$$(7.4) \quad \begin{aligned} \hat{V}(w_1, w_2) &= \hat{V}_{m_1, m_2}(w_1, w_2) \\ &= A + \gamma_1 B_1 w_1 + \gamma_2 B_2 w_2 + B_1 e^{-\gamma_1 w_1} + B_2 e^{-\gamma_2 w_2}, \end{aligned}$$

where  $\gamma_1 > 0, \gamma_2 > 0$  solve the quadratic equations

$$\left(\frac{\lambda_1}{\mu_1^2} + \frac{\lambda_2}{\mu_2^2}\right) \gamma_1^2 + b_1 \gamma_1 - 1 = 0, \quad \frac{\lambda_2}{\mu_3^2} \gamma_2^2 + b_2 \gamma_2 - 1 = 0,$$

and  $B_1 = c_1\mu_1/\gamma_1$ ,  $B_2 = c_3\mu_3/\gamma_2$ ,  $A = -\gamma_1 b_1 B_1 - \gamma_2 b_2 B_2$ .

The formula  $z_2^* = 0$  in (6.3) suggests that customer class 2 should always have priority, a fact already established by Chen, Yang & Yao (1991). Thus, for the queueing networks, we define the stationary control law (independent of  $n$  in this case)

$$Y(z) \triangleq \begin{cases} 1 & \text{if } z_1 > 0 \text{ or } z_2 > 0, \\ 0 & \text{if } z_1 = z_2 = 0, \end{cases}$$

$$U(z) \triangleq \begin{cases} 0 & \text{if } z_2 > 0, \\ 1 & \text{if } z_2 = 0. \end{cases}$$

One can show that  $(Y, U)$  is asymptotically optimal in the sense of (4.4) with  $\eta = 0$ . We omit the proof, focussing instead on the more complicated case IIA below.

## 8 Discussion of CASE II:

This is the case  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ . In Martins, Shreve & Soner (1993) the complete analysis is given only for

**CASE IIA:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 \leq 0$ ,  $c_2\mu_2 - c_3\mu_2 \geq 0$ ,  $c_2\mu_2 - c_1\mu_1 \geq 0$ .

In this case, the function  $\hat{h}$  given by (6.4) is nondecreasing in each variable separately. The optimal control processes for the workload problem are still given by (7.1), (7.2) and satisfy (7.3), but  $\hat{V}$  no longer has the simple closed form (7.4). Because

$$(8.1) \quad \hat{V}(w) = E \int_0^\infty e^{\alpha t} \hat{h}(W(t)) dt, \quad \forall w \in [0, \infty)^2,$$

the Feynman-Kac formula and elliptic regularity imply that  $\hat{V}$  is  $C^2$  on the open quadrant  $(0, \infty)^2$ ,  $\hat{V}$  is  $C^1$  on the closed quadrant  $[0, \infty)^2$ , and

$$(8.2) \quad \hat{V}_1(0, w_2) = \hat{V}_2(w_1, 0) = 0 \quad \forall (w_1, w_2) \in [0, \infty)^2,$$

$$(8.3) \quad \alpha \hat{V} - \hat{\mathcal{L}} \hat{V} - \hat{h} = 0 \text{ on } (0, \infty)^2,$$

where

$$\hat{\mathcal{L}} \hat{\varphi} \triangleq -b_1 \hat{\varphi}_1 - b_2 \hat{\varphi}_2 + \left( \frac{\lambda_1}{\mu_1^2} + \frac{\lambda_1}{\mu_2^2} \right) \hat{\varphi}_{11} + \frac{\lambda_2}{\mu_2 \mu_3} \hat{\varphi}_{12} + \frac{\lambda_2}{\mu_3^2} \hat{\varphi}_{22}$$

for any  $C^2$  function  $\hat{\varphi} : (0, \infty)^2 \rightarrow \mathbb{R}$ .

**CASE IIB:**  $c_1\mu_1 - c_2\mu_2 + c_3\mu_2 > 0$ ,  $c_2\mu_2 - c_3\mu_2 < 0$ ,  $c_2\mu_2 - c_1\mu_1 \geq 0$ .

Now  $\hat{h}$  is strictly decreasing in  $w_1$  for  $w_1 \in [0, \mu_3 w_2 / \mu_2]$ , which suggests that  $w_1$  should not be allowed to fall too far below  $\mu_3 w_2 / \mu_2$ . Numerical experimentation supports the conjecture that there is a continuous, increasing function  $\Psi_2 : [0, \infty) \rightarrow [0, \infty)$  such that the optimal control process  $m_1$  in the workload control problem acts whenever  $W_1(t) = \Psi_2(W_2(t))$  to ensure that the inequality  $W_1(t) \geq \Psi_2(W_2(t))$  is always satisfied. The rest of the conjecture was set out in Section 2.

**CASE IIC, IID:** The functions  $\Psi_1$  and  $\Psi_2$  appearing in the Section 2 conjectures about these cases are the free boundaries on which reflection should occur in the optimal control of the workload processes.

## References

- [1988] Barles, G. & Perthame, B., Exit time problems in optimal control and vanishing viscosity solutions of Hamilton-Jacobi equations, *SIAM J. Control Optimization* 26, 1133-1148.
- [1991] Chen, H., Yang, P. & Yao D., Control and scheduling in a two-station queueing network: optimal policies and heuristics, preprint.
- [1984] Crandall, M. Evans, L.C. & Lions, P. L., Some properties of viscosity solutions of Hamilton-Jacobi equations, *Trans. Amer. Math. Soc.* 282, 487-502.
- [1984] Crandall, M. & Lions, P. L., Viscosity solutions of Hamilton-Jacobi equations, *Trans. Amer. Math. Soc.* 277, 1-43.
- [1992] Crandall, M., Ishii, H. & Lions, P.L., User's guide to viscosity solutions of second order partial differential equations, *Bull. Amer. Math. Soc.* 27/1, 1-67.
- [1993] Fleming, W. & Soner, H. M., *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York.
- [1985] Harrison, J. M., *Brownian Motion and Stochastic Flow Systems*, Wiley, New York.
- [1988] Harrison, J. M., Brownian models of queueing networks with heterogeneous customer populations, In: *Stochastic Differential Systems, Stochastic Control Theory and Applications*, eds. W. Fleming & P. -L. Lions, IMA Vol. 10 (Springer-Verlag, New York), 147-186.
- [1989] Harrison, J. M. & Wein, L. M., Scheduling networks of queues: heavy traffic analysis of a simple open network, *Queueing Systems* 5, 265-280.
- [1990] Harrison, J. M. & Wein, L. H., Scheduling networks of queues: heavy traffic analysis of a two-station closed network, *Operations Research* 38, 1052-1064.
- [1993] Krichagina, E. V., Lou, S.X.C., Sethi, S. P., & Taksar, M. I., Production control in a failure-prone manufacturing system: diffusion approximation and asymptotic optimality, *Ann. Appl. Probab.* 3, 421-453.
- [1990] Kushner, H. J. & Martins, L. F., Routing and singular control for queueing networks in heavy traffic, *SIAM J. Control Optimization* 28, 1209-1233.



- [1991] Kushner, H. J. & Martins, L. F., Limit theorems for pathwise average cost per unit time problems for controlled queues in heavy traffic, preprint.
- [1988] Kushner, H. J. & Ramachandran, K. M., Nearly optimal singular controls for wideband noise driven systems, *SIAM J. Control Optimization* 26, 569-591.
- [1989] Kushner, H. J. & Ramachandran, K. M., Optimal and approximately optimal control policies for queues in heavy traffic, *SIAM J. Control Optimization* 27, 1293-1318.
- [1991] Laws, C. N., Resource pooling in queueing networks with dynamic routing, preprint.
- [1990] Laws, C. N. & Louth, G. M., Dynamic scheduling of a four-station queueing network, *Prob. Eng. Inf. Sci.* 4, 131-156.
- [1993] Martins, L.F., Shreve, S.E. & Soner, H.M., Heavy traffic convergence of a controlled, multi-class queueing system, forthcoming.
- [1990a] Wein, L. M., Optimal control of a two-station Brownian network, *Math. Operations Research* 15, 215-242.
- [1990b] Wein, L. M., Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs, *Operations Research* 38, 1065-1078.
- [1992] Wein, L. M., Scheduling networks of queues: heavy traffic analysis of a multistation network with controllable inputs, *Operations Research* 40, 5312-5334.

# STABILITY AND ERROR ESTIMATES OF STOCHASTIC INTEGRO-DIFFERENTIAL EQUATIONS WITH RANDOM PARAMETERS

G. S. Ladde

Department of Mathematics, The University of Texas at Arlington, Arlington, Texas 76019, U. S. A.

and

S. Sathananthan

Department of Mathematics and Center of Excellence in ISEM, Tennessee State University, Nashville, TN 37209, U.S.A.

## 1. Introduction.

Stochastic integro-differential equations arise in reactor dynamics, heat transfer, atomic scattering, population dynamics, and various other scientific disciplines. The randomness in the equation can be due to any possible combination of (i) random coefficients, (ii) random initial conditions, (iii) random forcing functions.

In this paper, utilizing a generalized variation of constants formula we have attempted to estimate the error between the solution of the stochastic and the solution of the deterministic version(mean) of a random integro-differential equation. Recently some attempts have been made in this direction for random differential equations [3,5], for random difference equations [6,7], and Itô-type stochastic differential and integro-differential equations [8,9,10].

We also have obtained some sufficient conditions to guarantee the stability of solutions. Furthermore, sufficient conditions are given for error estimates of solutions relative to corresponding smooth systems. The obtained results of the present study would provide a tool that verifies to what extent incorporating randomness in the system causes the change of behavior of the system relative to its deterministic version.

## 2. Preliminaries.

Let us consider the stochastic integro-differential equations(SIDE)

$$y' = f(t, y, \omega) + \int_{t_0}^t K(t, s, y(s, \omega)) ds, \quad y(t_0, \omega) = y_0(\omega) \quad (2.1)$$

---

This research reported herein was supported by the U. S. Army Research Office Grant No. DAAH04-93-G-0024.

and the corresponding deterministic initial value problem (DIVP)

$$m' = \hat{f}(t, m) + \int_{t_0}^t \hat{K}(t, s, m(s)) ds, \quad m(t_0) = E[y_0(\omega)] \quad (2.2)$$

which is obtained by ignoring the random disturbances in the system described by (2.1). In our subsequent analysis we will utilize the following random initial value problem (RIVP)

$$x' = \hat{f}(t, x) + \int_{t_0}^t \hat{K}(t, s, x(s)) ds, \quad x(t_0, \omega) = x_0(\omega) \quad (2.3)$$

In (2.1), (2.2) and (2.3),  $y, m, x \in R^n$ ,  $x_0, y_0 \in R[\Omega, R^n]$ ;  $E$  stands for the expected value of a random variable;  $(\Omega, F, P)$  is a complete probability space,  $f \in M[R_+ \times R^n, R[\Omega, R^n]]$ ,  $K \in M[R_+ \times R_+ \times R^n, R[\Omega, R^n]]$ , and  $f(t, y, \omega)$ ,  $K(t, s, y, \omega)$  are sample continuous in  $y$  for fixed  $t, s \in R^+$ ,  $\hat{f} \in C[R_+ \times R^n, R^n]$  and  $\hat{K} \in C[R_+ \times R_+ \times R^n, R^n]$ .

We assume that

(H<sub>1</sub>)  $f, K$  satisfies desired regularity conditions so that the initial value problem (2.1) has a sample solution process existing for  $t \geq t_0$ ;

(H<sub>2</sub>)  $\hat{f}_x, \hat{K}_x$  exist and  $\hat{f}_x \in C[R^+ \times R^n, R^{n \times n}]$ ,  $\hat{K}_x \in C[R^+ \times R^n \times R^n, R^{n \times n}]$ .

The above conditions implies that  $\hat{x}(t) = x(t, t_0, z_0)$  is unique solution process of (2.2) or (2.3) depending on the choice of  $z_0$ , where  $z_0$  is either  $y_0$  or  $x_0$ .

### 3. Variation of Constants Method.

In this section we present a generalized variation of constants method for stochastic systems of integro-differential equations with random coefficients. This method gives an integral representation of a function of a solution process of (2.1) with respect to the solution process of the random initial value problem (2.3) through  $(t_0, y_0(\omega))$ .

Theorem 3.1.[11] Let the hypotheses (H<sub>1</sub>), (H<sub>2</sub>) be satisfied, and  $y(t, \omega) = y(t, t_0, y_0(\omega), \omega)$  and  $x(t, \omega) = x(t, t_0, x_0(\omega))$  be the sample solution processes of (2.1) and (2.3) existing for  $t \geq t_0$  with  $x_0(\omega) = y_0(\omega)$ . Further assume that  $V_x(t, x, \omega)$  exists and is sample continuous for  $(t, x) \in R_+ \times R^n$ . Then,

$$V(t, y(t, \omega), \omega) = V(t_0, x(t, \omega), \omega) + \int_{t_0}^t \left\{ V_s(s, x(t, s, y(s, \omega)), \omega) \right. \\ \left. + V_x(s, x(t, s, y(s, \omega)), \omega) \Phi(t, s, y(s, \omega)) R(s, y(s, \omega), Ty(s, \omega), \omega) \right\} ds$$

$$+ \int_{t_0}^t \int_s^t [V_x(\sigma, x(t, \sigma, y(\sigma, \omega)), \omega) \Phi(t, \sigma, y(\sigma, \omega)) - V_x(s, x(t, s, y(s, \omega)), \omega) L(t, s, y(s, \omega), \omega)] \\ \cdot \hat{K}(\sigma, s, y(s, \omega)) d\sigma ds.$$

$$\text{where } R(t, y, Ty, \omega) = f(t, y, \omega) - \hat{f}(t, y) + \int_{t_0}^t [K(t, s, y(s, \omega), \omega) - \hat{K}(t, s, y(s, \omega))] ds. \quad (3.1)$$

**Proof:** From (2.1) and (2.3), system (2.1) can be rewritten as

$$y' = \hat{f}(t, y) + \int_{t_0}^t \hat{K}(t, s, y(s, \omega)) ds + R(t, y, Ty, \omega) \quad (3.2)$$

where  $R(t, y, Ty, \omega)$  is as defined in (3.1).

Let  $x(t, s, y(s, \omega))$  and  $x(t, \omega) = x(t, t_0, y_0(\omega))$  be the sample solution processes of (2.3) through  $(s, y(s, \omega))$  and  $(t_0, y_0(\omega))$ , respectively, and  $y(s, \omega) = y(s, t_0, y_0(\omega))$  be the sample solution process of (2.1) through  $(t_0, y_0(\omega))$ . Now we compute the total sample derivative of  $V(s, x(t, s, y(s, \omega)), \omega)$  with respect to  $s$  as

$$\begin{aligned} \frac{d}{ds} V(s, x(t, s, y(s, \omega)), \omega) &= V_s(s, x(t, s, y(s, \omega)), \omega) + V_x(s, x(t, s, y(s, \omega)), \omega) \left[ \frac{d}{ds} x(t, s, y(s, \omega)) \right] \\ &= V_s(s, x(t, s, y(s, \omega)), \omega) + V_x(s, x(t, s, y(s, \omega)), \omega) \left\{ -\Phi(t, s, y(s, \omega)) \hat{f}(s, y(s, \omega)) \right. \\ &\quad - \int_s^t L(t, \sigma; s, y(s, \omega)) \hat{K}(\sigma, s, y(s, \omega)) d\sigma + \Phi(t, s, y(s, \omega)) \left( \hat{f}(s, y(s, \omega)) + \int_{t_0}^s \hat{K}(s, \xi, y(\xi, \omega)) d\xi \right. \\ &\quad \left. \left. + R(s, y(s, \omega), Ty(s, \omega), \omega) \right) \right\} \\ &= V_s(s, x(t, s, y(s, \omega)), \omega) + V_x(s, x(t, s, y(s, \omega)), \omega) \left\{ - \int_s^t L(t, \sigma; s, y(s, \omega)) \hat{K}(\sigma, s, y(s, \omega)) d\sigma \right. \\ &\quad \left. + \Phi(t, s, y(s, \omega)) \left( \int_{t_0}^s \hat{K}(s, \xi, y(\xi, \omega)) d\xi + R(s, y(s, \omega), Ty(s, \omega), \omega) \right) \right\} \quad \text{w.p.1} \quad (3.3) \end{aligned}$$

Integrating in the sample sense both sides of (3.3) from  $t_0$  to  $t$ , and noting  $x(t, t, y(t, t_0, y_0(\omega), \omega)) = y(t, t_0, y_0(\omega), \omega)$ , we obtain,

$$\begin{aligned} V(t, y(t, \omega), \omega) &= V(t_0, x(t, \omega), \omega) + \int_{t_0}^t \left\{ V_s(s, x(t, s, y(s, \omega)), \omega) \right. \\ &\quad \left. + V_x(s, x(t, s, y(s, \omega)), \omega) \Phi(t, s, y(s, \omega)) R(s, y(s, \omega), Ty(s, \omega), \omega) \right\} ds \\ &\quad - \int_{t_0}^t V_x(s, x(t, s, y(s, \omega)), \omega) \int_s^t L(t, \sigma; s, y(s, \omega)) \hat{K}(\sigma, s, y(s, \omega)) d\sigma ds \end{aligned}$$

$$+ \int_{t_0}^t V_x(s, x(t, s, y(s, \omega)), \omega) \Phi(t, s, y(s, \omega)) \int_{t_0}^s \dot{K}(s, \xi, y(\xi, \omega)) d\xi ds \quad (3.4)$$

Using Fubini's Theorem the last term in (3.4) can be written as

$$\begin{aligned} & \int_{t_0}^t \int_{t_0}^s V_x(s, x(t, s, y(s, \omega)), \omega) \Phi(t, s, y(s, \omega)) \dot{K}(s, \xi, y(\xi, \omega)) d\xi ds. \\ &= \int_{t_0}^t \int_{t_0}^s V_x(\sigma, x(t, \sigma, y(\sigma, \omega)), \omega) \Phi(t, \sigma, y(\sigma, \omega)) \dot{K}(\sigma, s, y(s, \omega)) d\sigma ds. \end{aligned} \quad (3.5)$$

Using (3.5), (3.4) can be rearranged as (3.1) and hence the theorem follows.

**Corollary 3.1:** Let the assumptions of Theorem 3.1 holds. If  $V(t, x) = x$  with  $n=m$ , then (3.1)<sup>4</sup> reduces to

$$\begin{aligned} y(t, \omega) = x(t, \omega) &+ \int_{t_0}^t \Phi(t, s, y(s, \omega)) R(s, y(s, \omega), Ty(s, \omega), \omega) ds \\ &+ \int_{t_0}^t \int_s^t \left\{ \Phi(t, \sigma, y(\sigma, \omega)) - L(t, s, y(s, \omega)) \right\} \dot{K}(\sigma, s, y(s, \omega)) d\sigma ds. \end{aligned} \quad (3.6)$$

**Remark 3.1:** If  $f(t, y, \omega) = A(t, \omega)y$ ,  $K(t, s, y, \omega) = a(t, s, \omega)y$ , then (3.6) reduces to

$$y(t, \omega) = x(t, \omega) + \int_{t_0}^t \Phi(t, s) R(s, y(s, \omega), Ty(s, \omega), \omega) ds \quad (3.7)$$

**Corollary 3.2:** If  $V(t, x, \omega) = \|x\|^2$  then (3.1)<sup>4</sup> in Theorem 3.1 reduces to

$$\begin{aligned} \|y(t, \omega)\|^2 &= \|x(t, \omega)\|^2 + 2 \int_{t_0}^t x^T(t, s, y(s, \omega)) \Phi(t, s, y(s, \omega)) R(s, y(s, \omega), Ty(s, \omega), \omega) ds \\ &+ 2 \int_{t_0}^t \int_s^t \left\{ x^T(t, \sigma, y(\sigma, \omega)) \Phi(t, \sigma, y(\sigma, \omega)) \right. \\ &\quad \left. - x^T(t, s, y(s, \omega)) L(t, \sigma, y(s, \omega)) \right\} \dot{K}(\sigma, s, y(s, \omega)) d\sigma ds. \end{aligned} \quad (3.8)$$

**Theorem 3.2:** Suppose all the hypotheses of Theorem 3.1 hold. Then,

$$\begin{aligned} V(t, y(t, \omega) - \bar{x}(t), \omega) &= V(t_0, x(t, \omega) - \bar{x}(t), \omega) + \int_{t_0}^t \left\{ V_s(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)), \omega) \right. \\ &\quad \left. + V_x(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)), \omega) \Phi(t, s, y(s, \omega)) R(s, y(s, \omega), Ty(s, \omega), \omega) \right\} ds \\ &+ \int_{t_0}^t \int_s^t \left\{ V_x(\sigma, x(t, \sigma, y(\sigma, \omega)) - x(t, \sigma, \bar{x}(\sigma)), \omega) \Phi(t, \sigma, y(\sigma, \omega)) \right. \end{aligned}$$

$$-V_x(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)), \omega))L(t, s, y(s, \omega))\} \hat{K}(\sigma, s, y(s, \omega))d\sigma ds. \quad (3.9)$$

where  $\bar{x}(t) = x(t, t_0, z_0)$  is the solution process of either (2.2) or (2.3) depending on the choice of  $z_0$ .

**Proof:** By following the proof of Theorem 3.1, we have the relation,

$$\begin{aligned} \frac{d}{ds} V(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)), \omega) &= V_s(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)), \omega) \\ &+ V_x(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)), \omega) \frac{d}{ds} x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)) \end{aligned}$$

Using this and following the steps of Theorem 3.1, the theorem can be easily obtained.

**Corollary 3.3:** If  $V(t, x, \omega) = \|x\|^2$ , then (3.9) in Theorem 3.2 reduces to

$$\begin{aligned} \|y(t, t_0, y_0(\omega), \omega) - x(t, t_0, z_0)\|^2 &= \|x(t, t_0, y_0(\omega)) - x(t, t_0, z_0)\|^2 \\ &+ 2 \int_{t_0}^t \left( x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)) \right)^T \Phi(t, s, y(s, \omega)) R(s, y(s, \omega), Ty(s, \omega), \omega) ds \\ &+ 2 \int_{t_0}^t \int_s^t \left( x(t, \sigma, y(\sigma, \omega)) - x(t, \sigma, \bar{x}(\sigma)) \right)^T \Phi(t, s, y(s, \omega)) \\ &\quad - \left( x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s)) \right)^T L(t, \sigma; s, y(s, \omega)) \hat{K}(\sigma, s, y(s, \omega)) d\sigma ds. \end{aligned} \quad (3.10)$$

We recall that  $\bar{x}(t) = x(t, t_0, z_0)$  is the solution process of either (2.2) or (2.3) depending on the choice of  $z_0$ . In other words  $\bar{x}(t)$  is either  $m(t) = m(t, t_0, z_0) = x(t, t_0, m_0)$  or  $x(t, t_0, x_0(\omega))$ .

#### 4. Stability Analysis

By employing the preceding results, we give sufficient conditions for the  $p$ -th moment stability [4,8,9] of the trivial solution process of (2.1).

**Theorem 4.1:** Let the hypotheses  $(H_1)$ ,  $(H_2)$  be satisfied, and  $y(t, \omega) = y(t, t_0, y_0(\omega))$  and  $x(t, \omega) = x(t, t_0, x_0(\omega), \omega)$  be the sample solution processes of (2.1) and (2.3) existing for  $t \geq t_0$  with  $x_0(\omega) = y_0(\omega)$ ,  $V_x(t, x, \omega)$  exists and is sample continuous for  $(t, x) \in \mathbb{R}_+ \times \mathbb{R}^n$ . Furthermore,  $V(t, x)$ ,  $x(t, \omega)$ ,  $\Phi(t, s, y(s, \omega))$ ,  $R(s, y(s, \omega), Ty(s, \omega), \omega)$ ,  $\hat{f}(s, y(s, \omega))$ ,  $\hat{K}(t, s, y(s, \omega))$  satisfy

- (i)  $b(\|x\|^p) \leq \sum_{i=1}^m |V_i(t, x)| \leq a(\|x\|^p)$  for all  $(t, x) \in \mathbb{R}_+ \times \mathbb{R}^n$  where  $p \geq 1$ ,  $b \in \mathcal{V}\mathcal{K}$  and  $a \in \mathcal{C}\mathcal{K}$ ;
- (ii)  $\hat{f}(t, 0) \equiv 0$ ,  $\hat{K}(t, s, 0) \equiv 0$  with probability 1 for all  $s \leq t \in \mathbb{R}^+$ ;

$$(iii) \quad \|\mathfrak{D}V(s, x(t, s, y), \omega)\| \leq \lambda_1(t-s, \omega) \|V(s, y(s, \omega))\| + \int_{t_0}^s \lambda_2(s, \omega) \tilde{K}(t, \tau) \|V(\tau, y(\tau, \omega))\| d\tau$$

where  $t_0 < \tau \leq s \leq t$ , and  $\|y\|^p \leq \rho$ , where  $\rho$  is some positive real number,

$$\begin{aligned} \mathfrak{D}V(s, x(t, s, y)) &= V_s(s, x(t, s, y)) + V_x(s, x(t, s, y), \omega) \left\{ \Phi(t, s, y(s, \omega), \omega) R(s, y(s, \omega), Ty(s, \omega), \omega) \right. \\ &\quad \left. + \Phi(t, s, y(s, \omega), \omega) \int_{t_0}^s \tilde{K}(s, \xi, y(\xi, \omega)) d\xi - \int_s^t L(t, \sigma; s, y(s, \omega)) \tilde{K}(\sigma, s, y(s, \omega), \omega) ds \right\} \quad (4.1) \end{aligned}$$

$x(t, s, y)$  is the solution process of (2.3) through  $(s, y)$ ,  $\rho > 0$  and  $A \in C[R_+, R_+] \cap L^1[R_+, R_+]$  defined

by  $A(s, \omega) = \lambda_1(s, \omega) + \int_{t_0}^s \lambda_2(s, \omega) \tilde{K}(s, \tau, \omega) d\tau$ ,  $\lambda_1$  and  $\mu \in C[R_+, R_+] \cap L^1[R_+, R_+]$  and  $\mu$  is

defined by  $\mu = \int_{t_0}^t H_t(t, s, \omega) ds$  with  $H(t, s, \omega) = \lambda_1(s, \omega) + \int_s^t \tilde{K}(t, s, \omega) \lambda_2(\tau, \omega) d\tau$  such that

$$\frac{\partial H(t, s, \omega)}{\partial t} \geq 0 \quad \text{on } R_+ \times R_+ \text{ and } A(s, \omega) \text{ satisfies the relation}$$

$$E \left[ \exp \left[ \int_0^\infty A(s, t_0, \omega) ds \right] \right] \leq \exp \left[ \int_{t_0}^t \hat{A}(s) ds \right] \quad (4.2)$$

(iv)  $\|V(t_0, x(t, \omega), \omega)\| \leq \alpha(\|y_0(\omega)\|^p)$ , whenever  $E[\|y_0\|^p] \leq \rho$  for some  $\rho > 0$ , where  $\alpha \in \mathcal{C}\mathcal{K}$ . Then the trivial solution process of (2.1) is stable in the  $p$ -th moment.

**Theorem 4.2.** Assume that the hypotheses of Theorem 4.1 hold except that (iii) and (iv) are replaced by

$$(iii) \quad \|\mathfrak{D}V(s, x(t, s, y), \omega)\| \leq \lambda_1(s, \omega) \eta_1(t-s, \omega) \|V(s, y, \omega)\|$$

$$+ \lambda_2(s, \omega) \eta_2(t-s, \omega) \int_{t_0}^s \eta_3(s-\tau, \omega) \|V(\tau, y(\tau, \omega), \omega)\| \quad \text{for } t_0 \leq s \leq t,$$

$$(iv) \quad \|V(t_0, x(t, \omega), \omega)\| \leq \alpha(\|y_0(\omega)\|^p) \beta(t-t_0), \quad t \geq t_0,$$

provided  $E[\|y_0(\omega)\|^p] \leq \rho$  where  $\alpha \in \mathcal{C}\mathcal{K}$ ;  $\eta_1, \eta_2, \beta \in \mathcal{L}^1$ , and satisfies

$$\eta_1(t-s) \beta(s-t_0) \leq K \beta(t-t_0)$$

$$\eta_2(t-\tau) \eta_3(\tau-s) \leq K \eta_1(t-s) \lambda(t-s) \text{ for some } \lambda \in \mathcal{L}^1 \cap C[R_+, R_+] \text{ and}$$

$$\lim_{t \rightarrow \infty} \left[ \beta(t-t_0) \exp \left[ \int_{t_0}^t \hat{A}(s, t_0) ds \right] \right] = 0 \quad (4.3)$$

where  $A(s, t_0, \omega) = K\lambda_1(s, \omega) + K \int_{t_0}^s \lambda_2(s, \omega) \lambda(s-\xi, \omega) d\xi$ . Then the trivial solution process is

asymptotically stable in the  $p$ -th mean.

### 5. Error Estimate Results.

We present a few error estimate results by employing the method of variation of parameters with regard to systems of integro-differential equations with random coefficients.

**Theorem 5.1.** Let the assumptions of Theorem 3.2 be satisfied. Further assume that

$$(i) \quad b(\|x\|^p) \leq \sum_{i=1}^m |V_i(t, x, \omega)| \leq a(\|x\|^p),$$

$$(ii) \quad \sum_{i=1}^m |DV_i(s, x(t, s, y) - x(t, s, z), \omega)| \leq \lambda_1(t-s)C(\|y-z\|) + \lambda_2(t-s)C(\|z\|)$$

$$+ \int_{t_0}^s \lambda_3(t-s)\bar{a}(t, \tau)C(\|y(\tau) - z(\tau)\|^p) d\tau$$

where  $t_0 \leq \tau \leq s \leq t$ ,  $a \in \mathfrak{K}$  and it is differentiable,  $b \in \mathfrak{V}\mathfrak{K}$ ,  $C \in \mathfrak{K}$ , and  $DV_i(s, x(t, s, y) - x(t, s, z), \omega)$  is the  $i$ -th component of

$$DV(s, x(t, s, y) - x(t, s, z), \omega)$$

$$= V_s(s, x(t, s, y) - x(t, s, z), \omega) + V_x(s, x(t, s, y(s, \omega)) - x(t, s, \bar{x}(s))) \{ \Phi(t, s, y(\omega)) R(s, y, Ty, \omega) \}$$

$$+ \Phi(t, s, y(s, \omega)) \int_{t_0}^s \dot{K}(s, \xi, y(\xi, \omega)) d\xi - \int_s^t L(t, \sigma; s, y(s, \omega)) \dot{K}(\sigma, s, y(s, \omega)) ds \} \quad (5.1)$$

$p \geq 1$ ,  $\lambda_1, \lambda_2, \lambda_3 \in C[R_+, R_+] \cap L^1[R^+, R^+]$ . Let us define  $H(s)$ ,  $\frac{d}{ds} H = \frac{1}{h(s)}$ ,  $h(s) = C((b^{-1}(s))^{\frac{1}{p}})$  and assume that  $H \in \mathfrak{K}$ .

Then,

$$E[\|y(t, \omega) - \bar{x}(t)\|^p] \leq b^{-1}[r(t, \omega)] \quad (5.2)$$

for  $t \geq t_0$ , where  $y(t, \omega)$  and  $x(t, \omega)$  are the solution process of (2.1) and (2.3) through  $(t_0, y_0(\omega))$ , and  $\bar{x}(t) = x(t, t_0, z_0)$  is the solution process of (2.2) or (2.3) depending on the choice of  $z_0$ :  $\beta(s, \omega)$  is the absolute value of the sum of  $\lambda_2(0, \omega)C\|\bar{x}(s)\|^p$  and the time derivative of a  $(\|x(t, \omega) - \bar{x}(t)\|^p)$ ;  $r(t, \omega)$  is the maximal solution process of the integro-differential equation

$$m'(t, \omega) = \beta(t, \omega) + \lambda_1(0) h(m(t, \omega)) + \int_{t_0}^t \lambda_3(0, \omega) \bar{a}(t, \tau) h(m(s, \omega)) ds. \quad (5.3)$$

**Corollary 5.1:** Suppose that all the hypotheses of Theorem 5.1 are satisfied except the differentiability of  $a$  and assumption (ii) are replaced by

$$\|\Phi(t, s, y)\| \leq K \text{ for } t_0 \leq s \leq t, y \in R^n, \quad (5.4)$$

and assumption (ii) is valid whenever (5.4) holds, where  $K$  is a positive real number. Then (5.2)



reduces to

$$E[\|y(t, \omega) - \bar{x}(t)\|^p] \leq b^{-1} r(t, \omega)$$

where  $r(t, \omega)$  is the maximal solution process of the integro-differential equation

$$m'(t, \omega) = \lambda_2(0, \omega) C(\|\bar{x}(t)\|) + \lambda_1(0) h(m(t, \omega)) + \int_{t_0}^t \lambda_3(0, \omega) \bar{a}(t, \tau) h(m(s, \omega)) ds, \quad t \geq t_0 \quad (5.5)$$

The details of the proofs of these presented results will appear elsewhere.

#### References:

1. Bharucha-Reid, A. T., Random Integral Equations, Academic Press, New York, 1972.
2. Hu, S., Lakshmikantham, V. and Rama Mohana Rao, M., Nonlinear Variation of Parameters formula for Integro-differential equations of Volterra type, J. M. M. A., 29(1988), pp. 223-230
3. Ladde, G. S., Variational Comparison Theorem and Perturbations of Nonlinear systems, Proc. of AMS, Vol. 52, pp. 181-187, 1975.
4. Ladde, G. S. and Lakshmikantham, V., Random Differential Inequalities, Academic Press, New York 1980.
5. Ladde, G. S. and Sambandham, M., Error estimate of solutions and means of solutions of stochastic differential systems, Jour. of Math. Phys., Vol. 24 1983.
6. Ladde, G. S. and Sambandham, M., Variation of Constants formula and error estimates to stochastic difference equations, J. Math. and Phys. Sci., Vol. 22(1988), pp. 557-584.
7. Ladde, G. S. and Sambandham, M., Random difference inequalities, Trends in Theory and Practice of Non-linear Analysis (Ed. V. Lakshmikantham) Elsevier Science Publishers (North-Holland)(1985), 231-240.
8. Ladde, G. S., Sambandham, M. and Sathananthan, S., Comparison Theorem and its Applications, General Inequalities 6, (1992), pp. 321-342.
9. Ladde, G. S. and Sathananthan, S., Itô-type systems of stochastic integro-differential equations. Integral Methods in Science and Engineering-91, North-Holland Publishers(1991), pp. 75-89.
10. Ladde, G. S. and Sathananthan, S., Error estimates and stability of Itô-type systems of nonlinear stochastic integro-differential equations, Jour. Appl. Anal., Vol. 43, pp. 163-189, 1992.
11. Ladde, G. S. and Sathananthan, S., Stochastic Integro-Differential Equations with Random Parameters—I, Dynamic Systems and Applications, Vol. 1(1992), pp. 369-390.
12. Lakshmikantham, V., Leela, S. and Martynyuk, A. A., Stability Analysis of Nonlinear Systems, Marcel Dekker, New York and Basel, 1989.

# NUMERICAL TREATMENT OF RANDOM DIFFERENTIAL EQUATIONS

G. S. Ladde

Department of Mathematics, The University of Texas at Arlington, Arlington, TX 76019, U.S.A.

S. Sathananthan

Dept. of Math. and Center of Excell. in ISEM, Tennessee State University, Nashville, TN 37209, U.S.A.

and

R. Pirapakaran

Division of Natural and Computational Sciences, Wiley College, Marshall, TX 75670, U.S.A.

**Abstract:**—We considered several population models such as exponential growth, logistic growth and competition models with random coefficients and random initial conditions as random parameters. If, one or more of the initial conditions are degenerate constants, then Liouville's theorem, which describes the evolution of the Jacobian of the mapping is no longer applicable. Based on a numerical technique developed by Bellomo and Pistone [2], we obtained joint probability density function for the dependent variables or the marginal probability density function for the individual dependent variables. Numerical methods are also explored in these cases.

## 1. INTRODUCTION

Modern population biology is based on the fundamental models such as exponential growth, logistic growth, and competition models. Yet, their adequacy is questionable, in part, due to their deterministic properties. There has been a number of studies based on these models and treated the coefficients or initial conditions to reflect random environmental fluctuations [5]. Our analysis is based upon treating one or more, and perhaps all, of the initial data to be deterministic. In this approach, Liouville's theorem, which describes the evolution of the Jacobian of the mapping is no longer applicable. A more generalized valid technique, utilizing standard numerical methods for solving differential equations has been developed by Bellomo and Pistone [2] and utilized by Harlow and Delph [3] in solving differential equations numerically. This technique yields Liouville's theorem as a *special case* and utilizes the direct mapping between the nondegenerate random variables and the space of dependent variables. We have utilized this technique, in obtaining the joint density or the marginal density of the individual populations. The first and second moments are obtained to study the qualitative properties of populations. Standard numerical techniques are used for various cases.

The research reported herein was supported by the Army Research Office Grant No. DAAH-04-93-G-0024 and the National Security Agency Grant No. MDA-904-93-H-2002.

This paper was presented at the Tenth Conference in this series.

## 2. PRELIMINARIES

Consider the random differential equation,

$$\frac{dX}{dt} = g(t, X, \Lambda); \quad X(t_0) = X_0 \quad (2.1)$$

where  $X = (X_1, X_2, X_3, \dots, X_n)^T$ ,  $g = (g_1, g_2, \dots, g_n)^T$ , and  $T$  denotes transpose. The parameters  $X_0 = (X_{10}, X_{20}, \dots, X_{n0})^T$  with joint probability density function  $f_{X_0}(x_0)$  and  $\Lambda = (\Lambda_1, \Lambda_2, \dots, \Lambda_m)$  with joint probability density  $f_\Lambda(\lambda)$  are assumed to be random.  $\Lambda$  and  $X_0$  are assumed to be independent. We would like to consider the broader class of problems in which at least part of  $X_0$  is degenerate. We need the following results in our subsequent analysis.

**Lemma 2.1 [3]** Assume that (2.1) has a unique solution  $h(t, t_0, X_0)$  on  $[a, b]$ . Let  $f_X$  and  $f_\Lambda$  be continuous on the domain  $D(t, X, \Lambda)$  of  $(t, X, \Lambda)$  of dimension  $(n+m+1)$ . Then the solution is continuously differentiable on  $\{a < t < b; U\}$  where  $U = \{a < t_0 < b; |X_0 - \Psi(t_0)| + |\Lambda - \Lambda_0| < \delta\}$ .

Furthermore,  $\frac{\partial X}{\partial \Lambda_i} = \frac{\partial h}{\partial \Lambda_i} = Y$  is the solution of the initial value problem

$$\frac{dY}{dt} = f_X(t, h(t, t_0, X_0, \Lambda), \Lambda) Y + \frac{\partial f}{\partial \Lambda_i}(t, h(t, t_0, X_0, \Lambda), \Lambda); \quad Y(t_0) = 0 \quad (2.2)$$

Similarly,  $\frac{\partial h}{\partial X_{i0}} = Z$  satisfies

$$\frac{dZ}{dt} = f_X(t, h(t, t_0, X_0, \Lambda), \Lambda) Z; \quad Z(t_0) = e_i \quad (2.3)$$

where  $e_i$  is the vector  $e_i = (0, 0, 0, \dots, 1, 0, 0, \dots, 0)^T$ .

**Lemma 2.2 [4]** Let  $X_{00}$  represent the nondegenerate probability subspace contained in  $X_0$ , and the  $\dim(X_{00}) = \ell < n$ , where  $n = \dim X_0$ . We consider the following three different cases.

**Case(i):** We consider the situation, in which  $0 \leq \ell < n$ ,  $1 \leq m < n$ , and  $\ell + m = n$ . In this case,  $\dim[X_{00}, \Lambda] = \dim X$ . Therefore, the mapping  $(X_{00}, \Lambda) \rightarrow X = h(t, t_0, X_{00}, \Lambda)$  is well-defined. If the mapping is assumed to be continuous, and well-defined. If the mapping is assumed to be continuous, and one-one, then  $h^{-1}(t, t_0, X) = (X_{00}, \Lambda)$  exists, and the Jacobian  $J(t, t_0, X)$  of the inverse mapping, given by,  $h^{-1}: X \rightarrow (X_{00}, \Lambda)$  and

$$J(t, t_0, X) = \left| \frac{\partial h^{-1}}{\partial X} \right| = \left| \frac{\partial X_0}{\partial X} \right| \quad (2.4)$$

The probability density function for  $(X_{00}, \Lambda)$  to that of  $X$  is given by,

$$f_X(t, t_0, X) = f_{\Lambda, X_{00}}(t, t_0, h^{-1}(t, t_0, X)) \cdot |J(t, t_0, X)| \quad (2.5)$$

Case(ii): We consider the situation, in which  $0 \leq \ell < n$ ,  $1 \leq \mathcal{M} < n$ , and  $\ell + \mathcal{M} > n$ . Let  $q = \ell + \mathcal{M} - n$ . In order to obtain an invertible mapping, we carryout a standard augmentation of the probability space of  $X$ . To be specific, let us assume, without loss of generality, that  $q < \mathcal{M}$ , and define

$$B_1 = A_{\mathcal{M}-q+1}, B_2 = A_{\mathcal{M}-q+2}, B_3 = A_{\mathcal{M}-q+3}, \dots, B_q = A_{\mathcal{M}}.$$

We now consider the mapping  $(X_{00}, A) \rightarrow (X, B) = (h(t, t_0, X_0, A), B(A))$ ; which is well-defined. The Jacobian of the mapping is

$$K(t, t_0, A, X_{00}) = \left| \frac{\partial(h, B)}{\partial(A, X_{00})} \right|, \quad (2.6)$$

We assume that the mapping is invertible in the form  $(X_{00}, A) = (h^{-1}(t, t_0, X, B), A(B))$ , and hence, the Jacobian of the inverse mapping satisfies

$$\begin{aligned} J(t, t_0, X, B) &= \left| \frac{\partial(h^{-1}, A)}{\partial(X, B)} \right| = \frac{1}{K(t, t_0, (A, X_{00}))} \\ &= (h^{-1}(t, t_0, X, B), A) \end{aligned} \quad (2.7)$$

We can compute the joint probability density function on  $f_{X,B}$  from the relation

$$f_{X,B}(t, t_0, X, b) = f_{A, X_{00}}(t, t_0, h^{-1}(t, t_0, X, b), A(b)) |J(t, t_0, X, b)| \quad (2.8)$$

and the jpdf for  $X$  may be obtained from equation (2.8) by

$$f_X(t, t_0, X) = \int \dots \int f_{X,B}(t, t_0, X, b) db_1 db_2 db_3 \dots db_q \quad (2.9)$$

we will consider the case  $\ell + \mathcal{M} < n$ , the case for which the sum of the dimension of the nondegenerate initial conditions and the dimension of the random parameters  $A$  is strictly less than the dimension of  $X$ . In contrast to the first two cases of consideration the inverse transformation  $X \rightarrow (X_{00}, A)$  is not well-defined and hence the Jacobian  $J$  does not exist. The joint probability density function  $f_X$  may be either zero or does not exist. Nevertheless some, and possibly all the marginal pdf's may be constructed by means of the following procedure.

Case(iii): Let  $0 \leq \ell < n$ ,  $1 \leq \mathcal{M} < n$ , and  $\ell + \mathcal{M} < n$ . We will find the prf for the variable  $X_i = h_i(t, t_0, A, X_0)$ ,  $1 \leq i \leq n$ . We will assume that the solution for  $X_i$  is invertible in one of the random parameters  $(A, X_{00})$ . Without loss of generality, assume that this parameter is  $\lambda_1$ , so that we may write

$$\lambda_1 = h_i^{-1}(t, t_0, X_i, X_{001}, X_{002}, \dots, X_{00\ell}, \lambda_2, \lambda_3, \dots, \lambda_{\mathcal{M}}) \quad (2.10)$$

We perform the following augmentation, and define

$$B_1 = \lambda_2, \dots, B_{\mathcal{M}-1} = \lambda_{\mathcal{M}}, B_{\mathcal{M}} = X_{001}, \dots, B_{\mathcal{M}+\ell-1} = X_{00\ell} \quad (2.11)$$

By virtue of equation (2.11) the direct mapping  $(\Lambda, X_{00}) \rightarrow (X_i, B)$  exists, and the Jacobian has the simple form,

$$K(t, t_0, \Lambda, X_0) = \frac{\partial X_i}{\partial \lambda_i} \quad (2.12)$$

and hence the Jacobian for the inverse mapping satisfied,

$$J(t, t_0, X_i, B) = \frac{1}{K(t, t_0, h_i(t, t_0, X_i, \dots), \lambda_2 = B_1, \dots, X_{00l} = B_{\mathcal{M} + \mathcal{L} - 1})}$$

The joint pdf is given by,

$$f_{X_i, B}(t, t_0, X_i, b) = |J| f_{X_{00}, \Lambda}(t, t_0, \lambda_1 = h^{-1}, \lambda_2 = b_1, \dots, X_{00l} = b_{\mathcal{M} + \mathcal{L} - 1}) \quad (2.13)$$

for which the marginal pdf for  $X_i$  is given by

$$f_{X_i}(t, t_0, X_i) = \int \dots \int f_{X_i, B}(t, t_0, X_i, b) db_1 db_2 \dots db_{\mathcal{M} + \mathcal{L} - 1} \quad (2.14)$$

Now we are in a position to discuss the main results.

### 3. MAIN RESULTS

We will discuss the basic biological models such as exponential growth, logistic model and competition model treating parameters as random.

#### 3.1 EXPONENTIAL GROWTH MODEL:

Consider the exponential growth model of the form,

$$\frac{dN}{dt} = rN, \quad N(0) = N_0 \quad (3.1.1)$$

Here,  $r$  and  $N_0$  can be treated as random parameters.

##### 1. RANDOM INTRINSIC GROWTH RATE:

Assume that the initial condition  $N_0$  is deterministic constant and  $r$  is a random variable and independent of time  $t$ . Let  $r \sim \text{Normal}(\bar{r}, \sigma^2)$ .

Then,  $f_r(r) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{r-\bar{r}}{\sigma}\right)^2\right\}$ ,  $-\infty < r < \infty$

our aim is to study the mean and variance of  $\ln \frac{N}{N_0}$ . That is,  $E\left[\ln \frac{N}{N_0}\right]$  and  $V\left[\ln \frac{N}{N_0}\right]$ .

the solution of (3.1) is given by,  $N(t) = N_0 e^{rt}$  (3.1.2)

The space  $(X_{00}, \Lambda) = r$  and  $X = N$ . Therefore, the Jacobian of the mapping  $J: (X_{00}, \Lambda) \rightarrow X$  is defined and is given by,  $\left| \frac{\partial X}{\partial (\Lambda, X_{00})} \right| = \frac{\partial N}{\partial r} = Nt$ .

Therefore,  $f_N(N) = f_r\left(\frac{1}{t} \ln \frac{N}{N_0}\right) \frac{1}{Nt}$ ,  $-\infty < N < \infty$

$$= \frac{1}{\sqrt{2\pi\sigma}} \frac{1}{Nt} \exp - \frac{1}{2\sigma^2} \left\{ \frac{1}{t} \ln \frac{N}{N_0} - \bar{r} \right\}^2 \quad -\infty < \frac{1}{t} \ln \frac{N}{N_0} < \infty \quad (3.1.3)$$

Let,  $Z = \frac{1}{\sigma} \left( \frac{1}{t} \ln \frac{N}{N_0} - \bar{r} \right) \sim N(0,1)$  and we obtain,  $\ln \frac{N}{N_0} = t\sigma Z + t\bar{r}$ .

Therefore,  $E\left[\ln \frac{N}{N_0}\right] = t\bar{r}$  and  $V\left[\ln \frac{N}{N_0}\right] = t^2\sigma^2$ . (3.1.4)

If we assume that the initial condition  $N_0$  is a deterministic constant, and  $r$  is a random variable follows  $\text{Unif}(1,2)$  and independent of time. That is  $r \sim \text{Unif}(1,2)$ .

Then,  $f_N(N) = \frac{1}{tN}$  ;  $1 < \frac{1}{t} \ln \frac{N}{N_0} < 2$  (3.1.5)

Now,  $E\left[\ln \frac{N}{N_0}\right] = \int_{N_0 e^t}^{N_0 e^{2t}} \left(\ln \frac{N}{N_0}\right) \frac{1}{tN} dN = \frac{3}{2}t$  and  $E\left[\ln \frac{N}{N_0}\right]^2 = \int_{N_0 e^t}^{N_0 e^{2t}} \left(\ln \frac{N}{N_0}\right)^2 \frac{1}{tN} dN = \frac{7}{3}t^2$ ,

and hence we obtain,

$$V\left[\ln \frac{N}{N_0}\right] = \frac{1}{12}t^2 \quad (3.1.6)$$

## 2. RANDOM INITIAL CONDITION:

Assume that  $r$  is a deterministic constant, and  $N_0$  is a random variable and follows  $\exp(\lambda)$  and independent of time  $t$ .

That is,  $N_0 \sim \exp(\lambda)$ , and hence  $f_N(N) = e^{-rt} \cdot \lambda e^{-\lambda N} e^{-rt}$ ,  $N > 0$ .

Therefore,  $E[N] = \frac{1}{\lambda} e^{rt}$  and  $V[N] = \frac{1}{\lambda^2} e^{2rt}$ . (3.1.7)

If  $r$  is a deterministic constant and  $N_0$  is a random variable follows  $N(\mu, \sigma^2)$  then, we obtain  $E[N] = \mu e^{rt}$ , and  $V[N] = \sigma^2 e^{2rt}$ .

## 3. RANDOM INTRINSIC GROWTH RATE AND RANDOM INITIAL CONDITION:

Assume that both  $r$  and  $N_0$  are random variables and independent of time. In this case the Liouville's theorem fails and we use case (ii) of Lemma 2.2. Here,  $X = (N, r)$ , and the Jacobian of the direct mapping  $(X_{00}, A) \rightarrow X = (N, r)$  is given by,  $K = \left| \frac{\partial N}{\partial N_0} \right| = e^{rt}$ . If  $N_0, r$  are independently Uniformly(1,2) Then,

$$f_{N,r}(t, N, r) = e^{-rt}, \quad 1 < r < 2, \quad 1 < N^{-rt} < 2.$$

Then,

$$E[N] = \int_1^2 \int_{e^{-rt}}^{e^{2t}} N e^{-rt} dN = \frac{3}{2t} (e^{2t} - e^t), \quad E[N^2] = \frac{7}{6t} (e^{4t} - e^{2t}). \quad (3.1.8)$$

### 3.2 LOGISTIC GROWTH MODEL

Consider the Logistic-Growth model,

$$\frac{dN}{dt} = rN(1 - \frac{N}{K}), \quad N(0) = N_0 > 0 \quad (3.2.1)$$

Here  $r$ ,  $K$  and  $N_0$  can be treated as random parameters.

#### 1. RANDOM INTRINSIC GROWTH RATE:

Assume that  $r$  is a random variable while  $K$  and  $N_0$  are deterministic constants.

Let,  $r \sim \text{Normal}(\bar{r}, \sigma^2)$ .

Then,  $(X_{00}, A) = r \rightarrow X = N$  is well-defined and hence the Jacobian of the direct mapping  $= |\frac{\partial N}{\partial r}|$ .

The solution of (3.2.1) is given by,

$$N = \frac{KN_0 e^{rt}}{K - N_0 + N_0 e^{rt}} \text{ and Jacobian of the direct mapping} = \frac{(K - N_0)KN_0 e^{rt}}{(K + N_0 e^{rt} - N_0)^2 t}$$

$$\begin{aligned} \text{Therefore, } f_N(N) &= f_r(r \text{ in terms of } N, N_0 \text{ and } K) \cdot \frac{(K + N_0 e^{rt} - N_0)^2}{(K - N_0)KN_0 e^{rt} t} \\ &= \frac{1}{\sqrt{2\pi\sigma}} \frac{K}{\sigma(K - N)Nt} \exp - \frac{1}{2\sigma^2} \left[ \frac{1}{t} \ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} - \bar{r} \right]^2 \end{aligned} \quad (3.2.2)$$

Take,  $Z = \frac{1}{t\sigma} \ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} - \bar{r}$ , therefore  $\ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} = t\sigma Z + \bar{r}t$ , which gives,

$$E \left[ \ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} \right] = t\bar{r} \text{ and } V \left[ \ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} \right] = t^2 \sigma^2. \quad (3.2.3)$$

If  $r$  is uniform random variable  $\sim \text{Unif}(1,2)$  while  $K$  and  $N_0$  are deterministic constants.

$$\begin{aligned} \text{Then, } f_N(N) &= f_r \left[ \frac{1}{t} \ln \left\{ \frac{N}{K - N} \frac{K - N_0}{N_0} \right\} \right] \frac{K}{(K - N)Nt} \\ &= \frac{K}{(K - N)Nt}, \text{ whenever } 1 < \frac{1}{t} \ln \left\{ \frac{N}{K - N} \frac{K - N_0}{N_0} \right\} < 2 \end{aligned} \quad (3.2.4)$$

$$\text{Then, } E \left[ \ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} \right] = \frac{3}{2}t \text{ and } V \left[ \ln \left\{ \frac{N(K - N_0)}{N_0(K - N)} \right\} \right] = \frac{1}{12}t^2. \quad (3.2.5)$$

#### 2. RANDOM CARRYING CAPACITY:

Assume that  $K$  is a random variable,  $r$  and  $N_0$  are deterministic constants.

Let  $K \sim \text{Unif}(1,2)$ , then

$$f_N(t, N) = \frac{N_0(e^{rt} - 1)e^{rt}}{(N_0 e^{rt} - N)^2}; \quad 1 < \frac{NN_0(1 - e^{rt})}{N - N_0 e^{rt}} < 2$$

We can obtain the mean and variance as follows:

$$E[N(t)] = N_0^2(e^{rt}-1)e^{rt} \ln \left| \frac{N_0(e^{rt}-1)+1}{N_0(e^{rt}-1)+2} \right| + N_0 e^{rt}$$

$$\text{and } V[N(t)] = N_0^3(e^{rt}-1)e^{3rt} \frac{1}{[N_0(e^{rt}-1)+1][N_0(e^{rt}-1)+2]} \quad (3.2.6)$$

$$-N_0^4(e^{rt}-1)^2 e^{2rt} \ln \left| \frac{N_0(e^{rt}-1)+1}{N_0(e^{rt}-1)+2} \right|$$

### 3. RANDOM INTRINSIC GROWTH AND CARRYING CAPACITY:

Assume that  $r$  and  $K$  are random variables and  $N_0$  is a deterministic constant. In this case,  $(X_{00}, A) = (r, K)$ ,  $X = N$ . Therefore, the Liouville's theorem fails and we will consider the case (ii) of our Lemma 2.2 and augment the space  $X$ . Consider the mapping  $(X_{00}, A) = (r, K) \rightarrow X = (N, r)$  is well-defined and the Jacobian of the direct mapping,

$$K = \left| \frac{\partial X^T}{\partial (r, K)} \right| = \frac{(N_0 e^{rt} - N)^2}{N_0^2 (e^{rt} - 1) e^{rt}}, \text{ and hence the joint density of } N, r \text{ are given by}$$

$$f_{N,r}(t, N, r) = f_{r,K}(r, K \text{ in terms of } r, N, N_0) \cdot \frac{N_0^2 (e^{rt} - 1) e^{rt}}{(N_0 e^{rt} - N)^2} \quad (3.2.7)$$

Assume that  $r$  and  $K \sim \text{i.i.d. with Unif}(1, 2)$ . Then, the joint density of  $(N, r)$  given by (3.2.7) reduces to,

$$f_{N,r}(t, N, r) = \frac{N_0^2 (e^{rt} - 1) e^{rt}}{(N_0 e^{rt} - N)^2}; \quad 1 < r < 2, \quad 1 < \frac{NN_0(1-e^{rt})}{(N-N_0 e^{rt})} < 2. \quad (3.2.8)$$

The marginal density  $f_N(t, N)$  can be obtained by integrating (3.2.8) with respect to  $r$ .

### 4. RANDOM INTRINSIC GROWTH, CARRYING CAPACITY AND INITIAL CONDITION:

If the parameters  $r$ ,  $K$ , and  $N_0$  are random variables, i.i.d.  $\text{Unif}(1, 2)$ . Here Liouville's theorem fails and we need use case (ii) of the Lemma 2.2. We obtain, the joint density function  $f_{r,K,N}(t, r, K, N)$  as follows:

$$f_{r,K,N}(t, r, K, N) = \frac{K^2 e^{rt}}{[(K-N)e^{rt} + N]^2}, \quad 1 < r < 2, \quad 1 < K < 2, \quad \frac{Ke^{rt}}{(K-1)+e^{rt}} < N < \frac{2Ke^{rt}}{(K-2)+2e^{rt}} \quad (3.2.9)$$

The marginal density  $f_N(t, N)$  can be obtained by integrating (3.2.9) with respect to  $r$  and  $K$ .

### 3.3 COMPETING SPECIES MODEL

Consider the competing species model

$$x_i' = x_i(a_i - \sum_{j=1}^2 b_{ij}x_j), \quad i=1,2 \quad (3.3.1)$$



where  $x_i$  is the population density of the  $i$ -th species in the community. Let  $\alpha_1 > 0$ ,  $\alpha_2 > 0$  be the equilibrium states.

Then, (3.3.1) becomes,

$$w_i' = -b_{ii}w_i\alpha_i - \sum_{j=1}^2 b_{ij}(\alpha_i + w_i)w_j, \quad i=1,2 \quad (3.3.2)$$

The total number of parameters are  $b_{11}, b_{12}, b_{21}, b_{22}, w_1(0), w_2(0)$ . Assume that  $w_1(0), w_2(0)$  are deterministic constants,  $\alpha_1, \alpha_2$  are constants, and  $b_{11}, b_{12}, b_{21}, b_{22}$  are random. In this case  $(X_{00}, A) = (b_{11}, b_{12}, b_{21}, b_{22})$ ,  $X = (w_1, w_2)$  therefore, Liouville's theorem fails and we use case (ii) of Lemma 2.2 and we extend the space  $X$  as  $X = (w_1, w_2, b_{12}, b_{21})$ .

Therefore, the mapping  $(X_{00}, A) = (b_{11}, b_{12}, b_{21}, b_{22}) \rightarrow (w_1, w_2, b_{12}, b_{21})$  is well-defined and the Jacobian of the mapping

$$K = \frac{\partial X^T}{\partial (X_{00}, A)} = \det \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad (3.3.3)$$

$$\begin{aligned} \text{where } a_{11} &= \frac{\partial w_1}{\partial b_{11}}, \quad a_{12} = \frac{\partial w_2}{\partial b_{11}}, \quad a_{13} = \frac{\partial b_{12}}{\partial b_{11}}, \quad a_{14} = \frac{\partial b_{21}}{\partial b_{11}}, \quad a_{21} = \frac{\partial w_1}{\partial b_{12}}, \quad a_{22} = \frac{\partial w_2}{\partial b_{12}}, \\ a_{23} &= \frac{\partial b_{12}}{\partial b_{12}}, \quad a_{24} = \frac{\partial b_{21}}{\partial b_{12}}, \quad a_{31} = \frac{\partial w_1}{\partial b_{21}}, \quad a_{32} = \frac{\partial w_2}{\partial b_{21}}, \quad a_{33} = \frac{\partial b_{12}}{\partial b_{21}}, \quad a_{34} = \frac{\partial b_{21}}{\partial b_{21}}, \quad a_{41} = \frac{\partial b_{21}}{\partial b_{21}}, \\ a_{42} &= \frac{\partial w_2}{\partial b_{22}}, \quad a_{43} = \frac{\partial b_{12}}{\partial b_{22}}, \quad a_{44} = \frac{\partial b_{21}}{\partial b_{22}}. \end{aligned}$$

$$\text{By expanding the determinant, } K = \frac{\partial w_2}{\partial b_{11}} \frac{\partial w_1}{\partial b_{22}} - \frac{\partial w_1}{\partial b_{11}} \frac{\partial w_2}{\partial b_{22}}.$$

In order to obtain the Jacobian, let's use Lemma 2.1. Let's take  $w(t) = h(t, t_0, X_0, A)$  be the solution process of (3.3.2).

Then,  $\frac{\partial w}{\partial b_{ij}} = Y$  is the solution of the I.V.P.,

$$\frac{dY}{dt} = g_w(t, w(t), A)Y + \frac{\partial g(t, w(t), A)}{\partial b_{ij}}, \quad Y(t_0) = 0 \quad (3.3.4)$$

Similarly,  $\frac{\partial w}{\partial w_{i0}} = Z$  satisfies,

$$\frac{dZ}{dt} = g_w(t, w(t), A)Z, \quad Z(t_0) = e_i \quad (3.3.5)$$

From (3.3.2), (3.3.4) can be obtained as follows:

$$\frac{dY}{dt} = \begin{bmatrix} -b_{11}(\alpha_1 + 2w_1) - b_{12}w_2 & -b_{12}(\alpha_1 + w_1) \\ -b_{21}(\alpha_2 + w_2) & -b_{22}(\alpha_2 + 2w_2) - b_{21}w_1 \end{bmatrix} Y + \frac{\partial g}{\partial b_{ij}}(t, w(t), A), \quad Y(t_0) = 0 \quad (3.3.6)$$

$$\text{where } \frac{\partial g}{\partial b_{ij}}(t, w(t), A) = \begin{bmatrix} -(\alpha_1 + w_1)w_1 & -(\alpha_1 + w_1)w_2 & 0 & 0 \\ 0 & 0 & -(\alpha_2 + w_2)w_1 & -(\alpha_2 + w_2)w_2 \end{bmatrix},$$

We can solve (3.3.6) to obtain the Jacobian K.

Then, the Joint density of  $w_1, w_2, b_{12}, b_{21}$  can be obtained by using,

$$f_{w_1, w_2, b_{21}, b_{22}}(w_1, w_2, b_{21}, b_{22}) = f_{b_{11}, b_{12}, b_{21}, b_{22}}(b_{11}, b_{12}, b_{21}, b_{22}) \cdot \frac{1}{K} \quad (3.3.7)$$

Individual or marginal densities can be obtained by integrating (3.3.7).

#### 4. NUMERICAL METHODS

In this section, we will illustrate the analytical results obtained in the previous sections along with standard numerical methods to obtain the joint probability density functions and the marginal probability density functions of the solutions of exponential, logistic and competing species models involving random parameters and random initial conditions.

The numerical technique we have adapted is somewhat similar to [2], for the case in which the initial conditions are completely random. The method involves numerical computation of the inverse mapping  $X \rightarrow X_0$  and using standard numerical quadrature methods in order to obtain marginal probability density functions.

Example 4.1. Consider the exponential growth model,

$$\frac{dN}{dt} = rN, \quad N(0) = N_0 \quad (4.1)$$

Assume that the parameters  $N_0, r$  are random variables Uniformly distributed on the interval (1,2).

The solution of (4.1) can be easily obtained as,

$$N(t) = N_0 e^{rt}, \quad t \geq 0$$

and the joint probability density function of  $(N, r)$  are given by,

$$f_{N,r}(t, N, r) = e^{-rt}, \quad 1 < r < 2, \quad 1 < N e^{-rt} < 2 \quad (4.2)$$

The marginal probability density function of  $N$  in (4.2) can be obtained as follows:

$$f_N(t, N) = \begin{cases} \frac{1}{tN} [Ne^{-t} - 1], & e^t < N < e^{2t} \\ \frac{1}{tN} [Ne^{-t} - Ne^{-2t}], & e^{2t} < N < 2e^t \\ \frac{1}{tN} [2 - Ne^{-2t}], & 2e^t < N < 2e^{2t} \end{cases} \quad (4.3)$$

for  $t < \ln 2$ , and for  $t > \ln 2$

$$f_N(t, N) = \begin{cases} \frac{1}{tN} [Ne^{-t} - 1], & e^t < N < 2e^t \\ \frac{1}{tN}, & 2e^t < N < e^{2t} \\ \frac{1}{tN} [2 - Ne^{-2t}], & e^{2t} < N < 2e^{2t} \end{cases} \quad (4.4)$$

and  $f_N(t, N) \equiv 0$  outside the indicated region.

The Jacobian  $J$  of the inverse mapping  $N \rightarrow N_0$  is governed by,

$$\frac{dJ}{dt} = -rJ, \quad J(0) = 1 \quad (4.5)$$

In the numerical scheme proposed in [2] a value of  $r$  is chosen in the interval (1,2), along with a value of  $N$ . Then, (4.1) is integrated backwards in time from a given value of  $t$  to determine the corresponding value of  $N_0$ . This specifies the inverse mapping. Then, (4.5) is integrated forwards in time using the calculated time history  $N(t)$  to obtain  $J$ . This procedure will lead to the joint pdf  $f_{N,r}(t, r, N)$ . This joint pdf will be numerically integrated with respect to  $r$  to obtain the marginal pdf. The results are shown in Fig. 1. Again the numerical solution agrees with the exact solution to within the resolution of the plot.

When  $N_0$  is a deterministic constant, the pdf  $f_N(t, N)$  can be found to be

$$f_N(t, N) = \frac{1}{tN}, \quad N_0 e^t < N < N_0 e^{2t} \quad (4.6)$$

and identically equal to zero outside the indicated region.

Now the inverse mapping is the mapping  $r \rightarrow N$ . For a given value of  $t$ , this mapping may be calculated by a simple numerical shooting procedure which involves choosing a value of  $N$ , and then varying the value of  $r$  systematically so as to arrive at the chosen value of  $N$  when (4.1) is integrated forward in time. The Jacobian  $K$  of the mapping  $r \rightarrow N$  is just  $K = \frac{dN}{dr} = Y$ , and from (2.2)

$$\frac{dY}{dt} = rY + N, \quad Y(0) = 0 \quad (4.7)$$

In the numerical scheme, given the value of  $r$  from the shooting procedure and the time history  $N(t)$ , (4.7) can then be numerically integrated to obtain  $Y$ . The Jacobian of the inverse mapping  $N \rightarrow r$  is

given by  $J = 1/K=1/Y$ . An application of Lemma 2.2, equation (2.5) then yields directly the pdf for the solution  $f_N(t, N)$ .

We now consider an example from logistic growth model which illustrates the second of the three cases in Lemma 2.2.

Example 4.2. Consider the logistic growth model,

$$\frac{dN}{dt} = rN(K-N), \quad N(0) = N_0 > 0 \quad (4.8)$$

Here  $r, K$  are random parameters, and  $N_0$  is a deterministic constant.

In order to get the Jacobian, consider the mapping  $(X_{00}, A) = (r, K) \rightarrow X = (N, r)$  is well-defined and the Jacobian of the direct mapping,

$$K = \left| \frac{\partial(N, r)}{\partial(r, K)} \right| = \det \begin{bmatrix} \frac{\partial N}{\partial r} & \frac{\partial N}{\partial K} \\ \frac{\partial r}{\partial r} & \frac{\partial r}{\partial K} \end{bmatrix} = \frac{\partial N}{\partial K} = \frac{(N_0 e^{rt} - N)^2}{N_0^2 (e^{rt} - 1) e^{rt}}, \quad (4.9)$$

The joint density of  $N, r$  is given by,

$$f_{N,r}(t, N, r) = f_{r,K}(r, K \text{ in terms of } r, N, N_0) \cdot \frac{N_0^2 (e^{rt} - 1) e^{rt}}{(N_0 e^{rt} - N)^2} \quad (4.10)$$

Assume that  $r$  and  $K \sim$  i.i.d. with  $\text{Unif}(1, 2)$ . Then, the joint density of  $(N, r)$  given by (4.8) reduces to,

$$f_{N,r}(t, N, r) = \frac{N_0^2 (e^{rt} - 1) e^{rt}}{(N_0 e^{rt} - N)^2}; \quad 1 < r < 2, \quad 1 < \frac{NN_0(1 - e^{rt})}{(N - N_0 e^{rt})} < 2. \quad (4.11)$$

Therefore, the marginal density function of  $N$  can be obtained as follows:

$$f_N(t, N) = \begin{cases} \frac{1}{t} \ln \left[ \frac{N(N_0 - N)}{N_0 e^t - N} \right] + \frac{1}{tN} - \frac{(N - N_0)}{t(N_0 e^{2t} - N)}, & N_1 < N < N_2 \\ \frac{1}{t} \ln \left[ \frac{N_0 e^{2t} - N}{N_0 e^t - N} \right] + \frac{N_0 e^t (e^t - 1)(N - N_0)}{t(N_0 e^{2t} - N)(N_0 e^t - N)}, & N_2 < N < N_3 \\ \frac{1}{t} \ln \left[ \frac{(N_0 e^{2t} - N)(2 - N)}{(N - N_0)N} \right] + \frac{(2 - N)}{tN} - \frac{(N - N_0)}{t(N_0 e^{2t} - N)}, & N_3 < N < N_4 \end{cases}$$

$$\text{where } N_1 = \frac{N_0 e^{2t}}{N_0 (e^{2t} - 1) + 1}, \quad N_2 = \frac{N_0 e^t}{N_0 (e^t - 1) + 1}, \quad N_3 = \frac{2N_0 e^{2t}}{N_0 (e^{2t} - 1) + 2}, \quad N_4 = \frac{2N_0 e^t}{N_0 (e^t - 1) + 2}$$

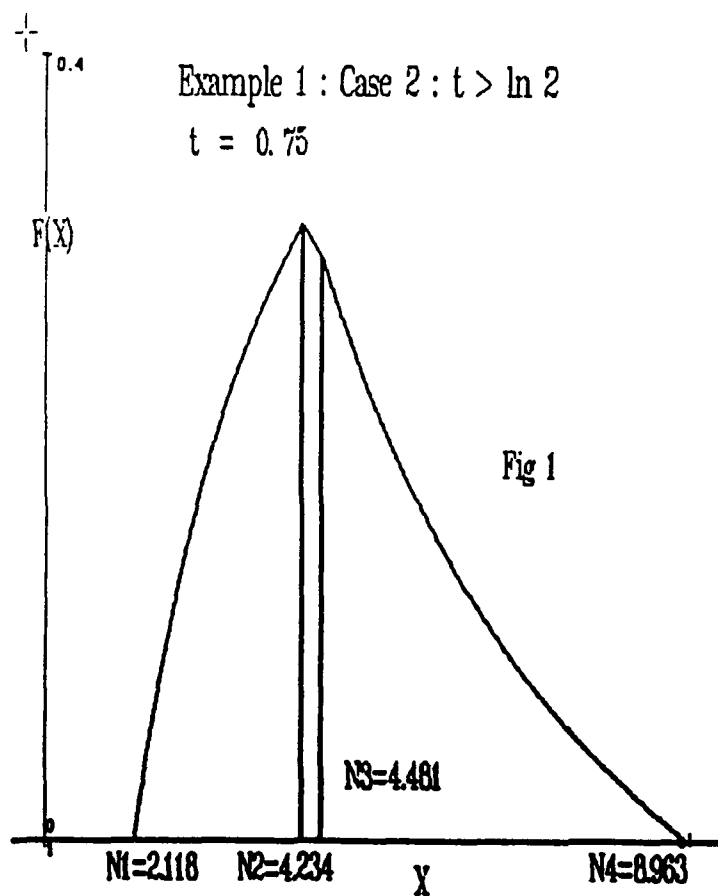
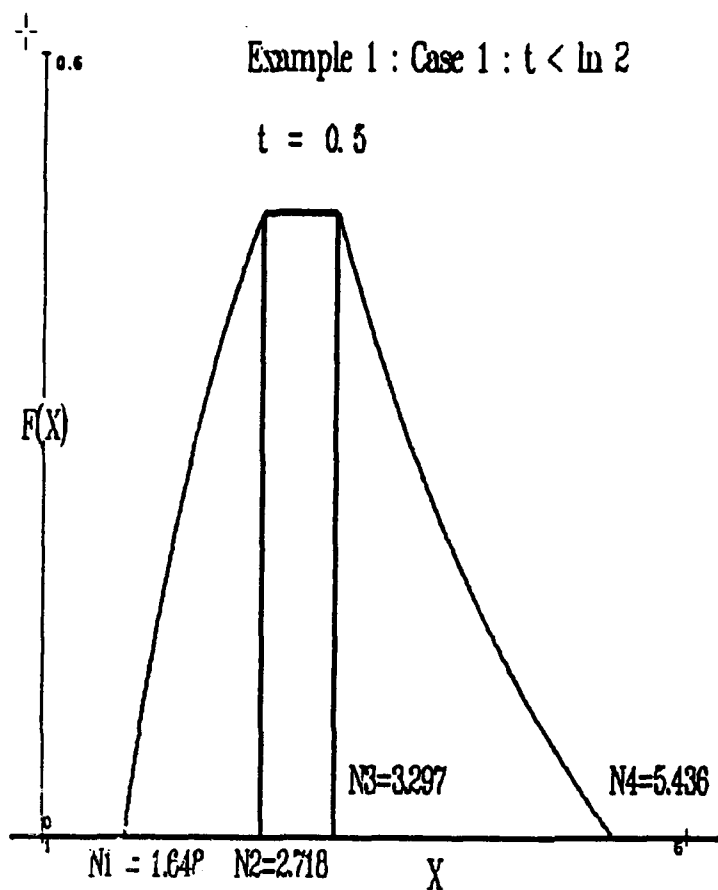


Fig 1

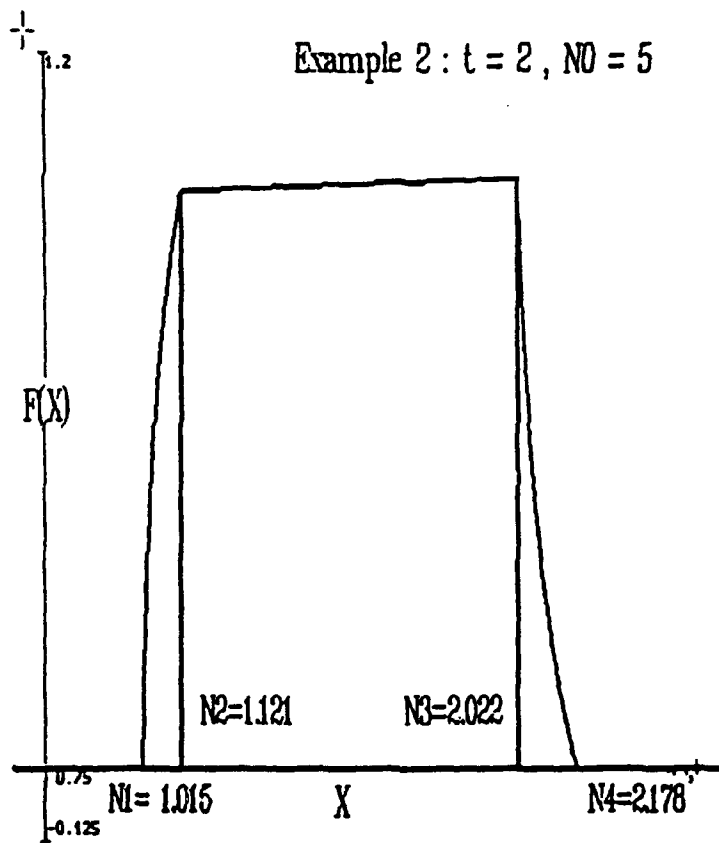
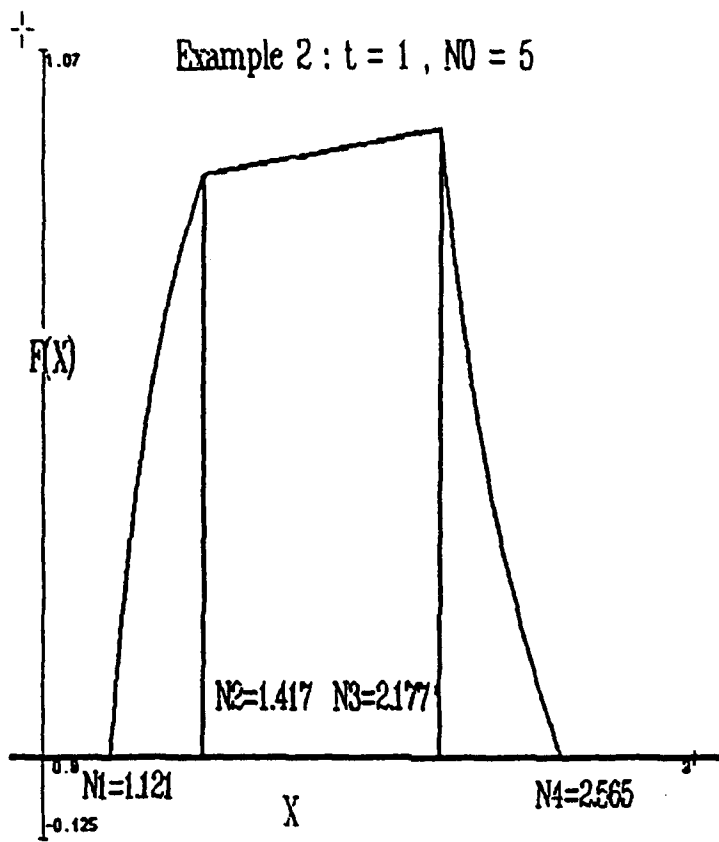


Fig 2

The numerical scheme is similar to the one described in Example 4.1. The results are shown in Fig. 2 for  $N_0 = 5$ . Again the numerical solution agrees with the exact solution to within the resolution of the plot.

#### REFERENCES:

1. Bellomo, N. and Pistone, G., Dynamical systems with a large number of degrees of freedom: a stochastic math. analysis for a class of determ. problems, Mech. Res. Comm. 6 (1979), 75-80.
2. Bellomo, N. and Pistone, G., Time evolution of the probability density under the action of a deterministic dynamical system, J. Math. Anal. Appl. 77(1980), 215-224.
3. Coddington, E. A. and Levinson, N., Theory of ordinary diff. equations, McGraw-Hill, 1955.
4. Harlow, D. G. and Delph, T.J., Num. Sol. of Random initial value problems, Math. and Comp. in Simulation, Vol.33, NO. 3, (1991), 243-258.
5. Soong, T. T., Random Diff. Equations in Sci. and Eng., Academic Press, NY, 1973.

# DIAGONALIZATION AND STABILITY OF TWO-TIME SCALE SINGULARLY PERTURBED LINEAR INTEGRO-DIFFERENTIAL SYSTEM

G. S. Ladde

Department of Mathematics, The University of Texas at Arlington, Arlington, TX 76019, U.S.A

and

S. Sathananthan

Dept. of Math. and Center of Excellence in Information Systems and Engineering

Tennessee State University, Nashville, TN 37209, U. S. A

## 1. INTRODUCTION

Singularly perturbed systems with two-time scale and, more generally, multi-time scale systems, often occur naturally due to the presence of small "parasitic" parameters multiplying derivatives [3,8,9]. In the last two decades the method of asymptotic expansion which is based on order reduction and boundary layer corrections has been widely used for such systems [7,9]. Recently, an alternative approach has been developed, where one develops a suitable non-singular linear transformation which partially or totally decouples the original system. The transformed system enables one to study the stability of original system with relative ease. This idea was initiated by Khalil and Kokotovic [2] for a two-time scale problem and Chang[1] for a general boundary value problem. Later, Ladde and Siljak[6], Ladde and Rajaluksmi [5], Ladde and Kathirkamanayagan[4] have used the idea for multiple time-scale and multi-parameter problems.

In this paper, a procedure to totally decouple a two-time scale singularly perturbed linear integro-differential system is developed. The procedure utilizes Chang's transformation [1] in a systematic and coherent manner. The fast and slow mode decomposition process provides a very elegant and powerful mechanism to investigate the stability and approximation analysis of the original system in terms of an auxiliary system corresponding to the decoupled system. Furthermore, the validity of the transformation is also discussed. The representation of the transformation in terms of the given coefficient matrices is given. Stability and the approximation to the solution of the original system are also investigated. Finally, an example illustrating the decoupling procedure and its applicability is presented.

---

This research reported herein was supported by the U. S. Army Research Office Grant No. DAAH04-93-G-0024 and the National Security Agency Grant No. MDA904-93-H-2002.

This paper was presented at the Tenth Conference in this series.



## 2. PRELIMINARIES

Consider a linear time-varying system of integro-differential equations

$$\mu \dot{x} = A x, \quad x(t_0) = x_0 \quad (2.1)$$

where,

$$A = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \quad \mu = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix}, \quad x = [x_1, x_2]^T; \quad x_i \in \mathbb{R}^{n_i} \text{ for } i=1,2 \text{ and } n = n_1 + n_2;$$

$$T_{11} = A_{11} + K_{11}, \quad T_{12} = A_{12} + K_{12}, \quad T_{21} = A_{21} + K_{21}, \quad T_{22} = A_{22} + K_{22};$$

$K_{ij}$ 's are linear Volterra operators and  $A_{ij}$ 's are matrices;  $\epsilon > 0$ .

We assume that

$H_1$ : The operators  $T_{ij}$ 's are bounded with respect to  $t \geq t_0$ ;

$H_2$ : Assume that  $T_{22}$  is invertible for all  $t \geq t_0$ ;

The  $\epsilon$ -boundary layer system can be obtained by putting  $\epsilon = 0$  in (2.1). It is given by,

$$\begin{aligned} \dot{x}_1 &= T_{11} x_1 + T_{12} x_2, \quad x_1(t_0) = x_{10} \\ \text{and } 0 &= T_{21} x_1 + T_{22} x_2, \quad x_2(t_0) = x_2^0 \neq x_{20} \end{aligned} \quad (2.2)$$

Let  $V$  and  $W$  are two Volterra operators defined as follows:

$$\begin{aligned} Vx &= D(t) x + \int_{t_0}^t a(t,s)x(s) ds \\ Wy &= E(t) y + \int_{t_0}^t b(t,s)y(s) ds \end{aligned} \quad (2.3)$$

where  $D, a$  are  $m \times n$  continuous matrices,  $E, b$  are continuous  $n \times m$  matrices and  $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ .

The composition of  $V$  and  $W$  are defined by:

$$(VW)(y) = V(Wy) = D(t)E(t)y + \int_{t_0}^t \left[ D(t) b(t,s) + a(t,s)E(s) + \int_s^t a(t,\theta)b(\theta,s)d\theta \right] y(s)ds.$$

and the derivative of the operator  $V$  is defined by,

$$(\dot{V}x)(t) = \dot{D}(t)x + a(t,t_0)x_0 + \int_{t_0}^t [a_t(t,s) + a_s(t,s)] x(s) ds.$$

The following two properties of the derivatives of Volterra operators  $V$  and  $W$  are needed in our subsequent discussion:

(1) The derivative of the Volterra operator  $V$  satisfies

$$\frac{d}{dt}[Vx(t)] = (\dot{V}x)(t) + (V\dot{x})(t) \quad \text{where } \dot{x} = \frac{d}{dt}x$$

(2) The derivative of the composition operator of V and W satisfies

$$\frac{d}{dt}[VWx(t)] = \dot{V}(Wx)(t) + V(\dot{W}x)(t).$$

### 3. DIAGONALIZATION PROCESS

In this section our prime objective is to develop a procedure to totally decouple the original system (2.1). This can be achieved by applying a transformation which decouples the fastest state variable in the coupled two-time scale system. The validity of such transformations will be discussed in the succeeding section. The following procedure briefly explains the method to totally decouple the original system (2.1).

We consider the following transformation  $S: C_1 \times C_2 \rightarrow C_1 \times C_2$  defined by

$$S = \begin{bmatrix} I_1 - \epsilon ML & -\epsilon M \\ L & I_2 \end{bmatrix} \quad (3.1)$$

where L and M are unknown linear Volterra operators which are functions of time;  $C_i = C[R, R^{n_i}]$  and  $I_1 - m \times m$  identity matrix,  $I_2 - n \times n$  identity matrix.

We note that the inverse of S is given by

$$S^{-1} = \begin{bmatrix} I_1 & \epsilon M \\ -L & I_2 - \epsilon LM \end{bmatrix} \quad (3.2)$$

Now, we can apply the transformation to the system (2.1) with  $m=n_1$ ,  $n=n_2$ :

$$\begin{aligned} Z &= SX \\ \text{So, } \dot{Z} &= (\dot{S}S^{-1} + S\mu^{-1}AS^{-1})Z \end{aligned} \quad (3.3)$$

Here,  $Z = (z_1, z_2)^T$ ,  $z_i \in R^{n_i}$  for  $i=1,2$ ;

$$\text{Set, } \dot{S}S^{-1} + S\mu^{-1}AS^{-1} = P, \text{ where } P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}.$$

Choose L and M in (3.1) so that  $P_{12} \equiv 0$  and  $P_{21} \equiv 0$ . These two identities leads to the following:

$$P_{12} \equiv 0 \text{ implies that, } \epsilon \dot{L} = T_{22}L - \epsilon LT_{11} + \epsilon LT_{12}L - T_{21} \quad (3.4)$$

$$P_{21} \equiv 0 \text{ implies that, } \epsilon \dot{M} = -MT_{22} + \epsilon(T_{11} - T_{12}L)M - \epsilon MLT_{12} + T_{12} \quad (3.5)$$

Under these conditions (3.4) and (3.5),  $P_{11}$  and  $P_{12}$  can be written as:

$$P_{11} = (T_{11} - T_{12}L) \text{ and } P_{22} = LT_{12} + \epsilon^{-1}T_{22}. \quad (3.6)$$

Hence the system (3.3) reduces to:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12}L \\ 0 & \epsilon^{-1}T_{22} + LT_{12} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}; \quad z(t_0) = z_0 \quad (3.7)$$

#### 4. VALIDITY AND APPROXIMATION OF THE TRANSFORMATIONS

In order to establish the validity of the transformation, we will establish the existence, uniqueness, boundedness and other fundamental properties of the following abstract Cauchy problem in  $L$  and  $M$ .

$$\begin{aligned} \epsilon \dot{L} &= T_{22}L + \epsilon L(T_{12}L - T_{11}) - T_{21} \\ \epsilon \dot{M} &= -MT_{22} + \epsilon(T_{22} - T_{12}L)M - \epsilon MLT_{12} + T_{12} \end{aligned} \quad (4.1)$$

with initial conditions,

$$\begin{aligned} L(t_0) &= T_{22}^{-1}(t_0) T_{21}(t_0) \\ M(t_0) &= T_{12}(t_0) T_{22}^{-1}(t_0). \end{aligned}$$

From continuity of the matrices and kernels in (2.1) and continuous differentiability of the right hand side of (2.1) relative to  $L$  and  $M$ , the existence and uniqueness of the problem (4.1) follows immediately.

**Remark 4.1:** The sufficient conditions to establish the inverse of  $T_{22}$  can be given as follows:

$$\text{Define, } T_{22}u = A_{22}(t)u + \int_{t_0}^t K_{22}(t,s)u(s)ds$$

$$\text{and, } T_{22}^{-1}u = C(t)u + \int_0^\infty F(t,s)u(s)ds$$

Then,

(i)  $C(t) = A_{22}^{-1}(t)$  and  $F(t,s)$  satisfies the integral equations

$$(ii) \quad A_{22}(t)F(t,s) + K_{22}(t,s)C(s) + \int_s^t K_{22}(t,\zeta)F(\zeta,s)d\zeta \equiv 0 \text{ and}$$

$$(iii) \quad C(t)K_{22}(t,s) + F(t,s)A_{22}(s) \int_s^t F(t,\zeta)K_{22}(\zeta,s)d\zeta \equiv 0, \quad (4.2)$$

We need the following assumptions in order to establish the results.

(H<sub>3</sub>):  $T_{22}^{-1}T_{21}$ ,  $T_{12}T_{22}^{-1}$  and their derivatives are bounded on  $[t_0, \infty)$ .

(H<sub>4</sub>):  $\nu(T_{22}) \leq -\alpha$ ,  $\alpha > 0$  where  $\nu(T_{22}) = \limsup_{h \rightarrow 0^+} \left[ \frac{\|I + hT_{22}\| - 1}{h} \right]$  is the logarithmic norm of

the linear Volterra operator  $T_{22}$ .

**Theorem 4.1:** Under the assumptions  $(H_1)-(H_4)$ , the abstract Cauchy problem (4.1) has almost one solution  $(L, M)$  existing on  $[t_0, \infty)$ .

Moreover,

$$\begin{aligned} L(t) &= \dot{L}(t) + 0(\epsilon) \\ M(t) &= \dot{M}(t) + 0(\epsilon) \end{aligned} \quad (4.3)$$

where

$$\dot{L} = -T_{22}^{-1}T_{21} \text{ and } \dot{M} = T_{12}T_{22}^{-1}.$$

**Proof:** The proof of the theorem can be formulated analogous to the result in [5] with certain modifications.

## 5. STABILITY RESULTS

In this section, we establish the main result concerning the approximate solution and the stability of the original system (2.1). An approximate solution of (2.1) can be obtained as follows:

The totally decoupled system of (2.1) can be written as,

$$\begin{aligned} \dot{u}_1 &= (T_{11} - T_{12}L)u_1 + 0(\epsilon), \quad u_1(t_0) = u_1^0 \\ \text{and} \quad \dot{u}_2 &= (\epsilon LT_{12} + T_{22})u_2 + 0(\epsilon), \quad u_2(t_0) = u_2^0 \end{aligned} \quad (5.1)$$

We note that system (5.1) can be considered as a perturbed system [6] of

$$\begin{aligned} \dot{v}_1 &= (T_{11} - T_{12}L)v_1 + 0(\epsilon), \quad v_1(t_0) = v_1^0 \\ \text{and} \quad \epsilon \dot{v}_2 &= (\epsilon LT_{12} + T_{22})v_2 + 0(\epsilon), \quad v_2(t_0) = v_2^0 \end{aligned} \quad (5.2)$$

where  $v^0 = \hat{T}^2 x_0$ .

This system can be considered as an auxiliary system of (5.1). We need an additional assumption to establish our results.

$$(H_5): \quad \limsup_{t \rightarrow \infty} \left[ \frac{1}{t-t_0} \int_{t_0}^t \nu(T_{11} - T_{12}L)(s) ds \right] \leq -\alpha_1 < 0.$$

**Lemma 5.1:** Let the assumptions  $H_5$ , and  $H_4$  hold. Then, one can choose  $\epsilon^* > 0$  such that for all  $\epsilon \leq \epsilon^*$ ,

- (i) The trivial solution of (5.1) is exponentially asymptotically stable.
- (ii)  $u_1 = v_1 + 0(\epsilon)$ ,  $u_2 = v_2 + 0(\epsilon)$ .

**Proof:** The proof of the lemma can be formulated analogous to the result in [6] with certain modifications.

**Theorem 5.1:** Under the assumptions of Theorem 4.1 and Lemma 5.1,

- (i) The trivial solution of (2.1) is exponentially asymptotically stable.
- (ii) The solution of (2.1) can be approximated by  $\hat{S}^{-1}z$ , where  $z$  is the solution of (5.2), that is  $x(t) = \hat{S}^{-1}z + o(\epsilon)$ , where

$$S^{-1} = \begin{bmatrix} I_1 & -\epsilon \hat{M} \\ \hat{L} & I_2 - \epsilon \hat{L} \hat{M} \end{bmatrix} \quad (5.3)$$

**Proof:** The proof of the lemma can be formulated analogous to the result in [4] with certain modifications.

**Example 5.1:** Consider the following system,

$$\begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \quad (5.4)$$

$$\text{where } T_{11}x = \frac{1}{2}x + \int_{t_0}^t -9e^{-7(t-s)}x(s) ds, \quad T_{12}x = -5x + \int_{t_0}^t 3e^{-7(t-s)}x(s) ds,$$

$$T_{21}x = -3x + \int_{t_0}^t -2e^{-5(t-s)}x(s) ds, \quad T_{22}x = -5x + \int_{t_0}^t -2e^{-2(t-s)}x(s) ds.$$

We obtain the following :

- (i) The operator  $T_{22}$  is invertible and the inverse operator of  $T_{22}$ ,

$$T_{22}^{-1}u = -\frac{1}{5}u + \int_{t_0}^t \frac{2}{25}e^{-\frac{12}{5}(t-s)}u(s) ds,$$

- (ii) The Logarithmic Norm of the transformation  $T_{22}$  is given by,

$$\mu(T_{22}) \leq -3.$$

- (iii) The solution of the Cauchy problem (4.1) is given by,

$$L = \hat{L} + o(\epsilon)$$

$$M = \hat{M} + o(\epsilon), \quad \text{where } \hat{L} \text{ and } \hat{M} \text{ are given by,}$$

$$\dot{L}u = \frac{3}{5}u + \int_{t_0}^t \left[ -\frac{9}{13}e^{-5(t-s)} - \frac{48}{325}e^{-\frac{12}{5}(t-s)} \right] u(s) ds,$$

and 
$$\dot{M}u = u + \int_{t_0}^t \left[ -\frac{4}{13}e^{-\frac{12}{5}(t-s)} - \frac{9}{13}e^{-5(t-s)} \right] u(s) ds.$$

(iv) The totally decoupled system can be obtained as,

$$u_1 = -\frac{5}{2}u + \int_{t_0}^t \left[ -\frac{117}{26}e^{-5(t-s)} - \frac{192}{199}e^{-\frac{12}{5}(t-s)} - \frac{16711}{2990}e^{-7(t-s)} \right] u(s) ds + o(\epsilon)$$

and 
$$\epsilon u_2 = -5u_2 + \int_{t_0}^t -2e^{-5(t-s)} u_2(s) ds + o(\epsilon) \quad (5.5)$$

(v) The system in (iv) can be considered as a perturbed system of,

$$v_1 = -\frac{5}{2}v_1 + \int_{t_0}^t \left[ -\frac{117}{26}e^{-5(t-s)} - \frac{192}{199}e^{-\frac{12}{5}(t-s)} - \frac{16711}{2990}e^{-7(t-s)} \right] v_1(s) ds$$

and 
$$v_2 = -5v_2 + \int_{t_0}^t -2e^{-2(t-s)} v_2(s) ds \quad (5.6)$$

with  $v_1(t_0) = v_1^0$ , and  $v_2(t_0) = v_2^0$ .

(vi) The solution of the original system by applying the main results can be approximated by,

$$\begin{aligned} x_1(t) &= v_1(t) + o(\epsilon) \\ x_2(t) &= -\dot{L}v_1(t) + v_2(t) + o(\epsilon) \end{aligned} \quad (5.7)$$

where  $v_1(t)$ ,  $v_2(t)$  are the solutions of (v).

Note that the Logarithmic Norm of  $\tau_a$ ,  $\nu(\tau_a) \leq -0.304$ , where

$$\tau_a u = -\frac{5}{2}u + \int_{t_0}^t \left[ -\frac{117}{26}e^{-5(t-s)} - \frac{192}{199}e^{-\frac{12}{5}(t-s)} - \frac{16711}{2990}e^{-7(t-s)} \right] u(s) ds$$

implies that the trivial solution of the original system (5.4) is exponentially asymptotically stable.

The details of the proofs of these presented results will appear elsewhere.

#### REFERENCES

1. Chang, K. W., Singular perturbations of a general boundary value problem, SIAM J. Math. Anal. 3(1972), pp. 520-526.
2. Khalil, H. K., and Kokotovic, P.V., D-stability and multi-parameter singular perturbations, SIAM J. Control Optim. 17(1979), pp. 55-65.

3. Kokotovic, P. V., Khalil, H. K. and O' Reily J., " Singular Perturbation Methods in Control: Analysis and Design", Academic Press, New York/London, 1986.
4. Ladde, G.S., and Kathirkamanayagan, M., Diagonalization and stability of large-scale singularly perturbed linear system, J. of Math. Anal. and Appl., Vol. 135, No. 1, 1988, pp. 38-60.
5. Ladde, G. S., and Rajaluksmi, S.G., Diagonalization and stability of multi-time scale singularly perturbed linear system, Appl. Math. and Comp. 16(1985), pp. 115-140.
6. Ladde, G.S., and Siljak, D. D., Multiparameter singular perturbations of linear systems with multiple time scales, Automatica 19(1983), pp. 385-394.
7. O' Malley, R. E., On initial value problems for nonlinear systems of differential equations with two small parameters, Arch. Rat. Mech. Anal. 40(1971), pp. 209-222.
8. Saksena, V. R., O' Reiley, J. and Kokotovic, P. V., Singular perturbations and time-scale methods in control theory: Survey 1976-83, Automatica 20(1984), pp. 273-293.
9. Smith, R., Singular Perturbation Theory: An Introduction with Applications", Cambridge University Press, New York, 1985.

**Eleventh Army Conference on  
Applied Mathematics and Computing**

**June 8-10, 1993**

**Participant List**

**Dr. Gerald R. Anderson**  
Associate Director  
USARO  
Mathematical & Computer Sciences Division  
(919) 549-4253  
e-mail: Jerry@BRL.Army.Mil

**Prof. Peter Beckman**  
Professor  
Bryn Mawr College  
Physics Department  
(215) 526-5361  
e-mail:

**Dr. Neal E. Blackwell**  
Engineer  
US. Army Belvoir R&D Center  
Environmental Equipment Development Team  
(703) 704-3899  
e-mail: nblackwell@belvoir.emg4.army.mil

**Dr. Bolindra N. Borah**  
Professor of Applied Mathematics  
NC. A&T University  
Department of Mathematics  
(929) 334-7822  
e-mail: borah@garfield.ncat.edu

**Prof. Deborah Brandon**  
Professor  
Carnegie Mellon University  
Department of Mathematics  
(412) 268-2545  
e-mail: Brandon@andrew.cmu.edu

**Dr. Aivars Celmins**  
Army Research Laboratory  
ACISD  
(410) 278-6986  
e-mail: celmis@brl.army.mil



**Dr. Jagdish Chandra**  
Director  
USARO  
Mathematical & Computer Sciences Division  
(919)549-4254  
e-mail: Chandra@ARO-emhl.army.mil

**Dr. Pao-Liu Chow**  
Professor  
Wayne State University  
Department of Mathematics  
(313)577-3197  
e-mail: plchow@waynest1.bitnet

**Dr. Kenneth Clark**  
U.S. Army Research Office  
Mathematical & Computer Sciences Division  
(919)549-4256  
e-mail: clark@adm.csc.ncsu.edu

**Dr. Joseph Michael Coyle**  
Benet Labs  
Research Division  
(518)266-5883  
e-mail:

**Mr. Terence M. Cronin**  
Computer Scientist  
Army IEWD  
AMSEL-RD-IEW-TRF  
(703)349-6939  
e-mail:

**Dr. Ben Cummings**  
Army Research Lab  
ACISD  
(410)278-6666  
e-mail: Cummings@BRL.MIL

**Prof. Rick Durrett**  
Professor  
Cornell University  
Department of Mathematics  
(607)255-8282  
e-mail: rtd@CornellA

**Prof. Nabil H. Farhat**  
**Professor**  
**University of Pennsylvania**  
**Department of Electrical Engineering**  
**(215)898-5882**  
**e-mail: farhat@ee.upenn.edu**

**Dr. Donald A. French**  
**Assistant Professor**  
**University of Cincinnati**  
**Mathematical Sciences Department**  
**(513)556-4039**  
**e-mail: French@ucunix.san.uc.edu**

**Dr. Avner Friedman**  
**Director**  
**University of Minnesota**  
**Institute for Mathematics and its Applications**  
**(612)624-6066**  
**e-mail: friedman@ima.umn.edu**

**Dr. James Glimin**  
**Professor**  
**State University of New York at Stony Brook**  
**Applied Mathematics & Statistics**  
**(516)632-8355**  
**e-mail: glimin@ams.sunysb.edu**

**LTC Robert P. Gocke, Jr.**  
**Research Fellow**  
**Defense Systems Management College**  
**Department of Research & Information**  
**(703)805-2525**  
**e-mail:**

**Prof. John W. Grove**  
**Professor**  
**State University of New York at Stony Brook**  
**Department of Applied Mathematics**  
**(516)632-8375**  
**e-mail: grove@ams.sunysb.edu**

**Dr. Aaron D. Gupta**  
**Mechanical Engineer**  
**AMSRL-WT-TD, Weapons Technology Directorate**  
**(410)278-6026**  
**e-mail: dasgupta@brl.mil**

**Dr. Morton E. Gurtin**  
Director  
Carnegie Mellon University  
Center for Nonlinear Analysis  
(412)268-3554  
e-mail: mg0c@andrew.cmu.edu

**Dr. Andrew Harrell**  
U.S. Army Engineer Waterways Experiment Station  
Geotechnical Laboratory  
(601)634-3382  
e-mail: h3gm0ah0@wes.army.mil

**Dr. Jeffery P. Holland**  
Director  
USAE Waterways Experiment Station  
Computational Hydraulics Institute  
(601)634-2644  
e-mail: holland@hr1.wes.army.mil

**Prof. William J. Hrusa**  
Professor  
Carnegie Mellon University  
(412)268-2545  
e-mail:

**Dr. William Jackson**  
U.S. Army Tank-Automotive Command  
(313)574-7530  
e-mail: Jacksonw@tacom-emh165.army.mil

**Prof. David Kinderlehrer**  
Professor  
Carnegie Mellon University  
(412)268-2545  
e-mail: dk3l@andrew.cmu.edu

**Dr. Petr Kloucek**  
Research Associate  
University of Minnesota  
School of Mathematics  
(612)625-0072  
e-mail: kloucek@math.umn.edu

**Dr. P.S. Krishnaprasad**  
Professor of Electrical Engineering

University of Maryland  
Electrical Engineering & Institute for Systems Research  
(301)405-6843  
e-mail: krishna@src.umd.edu

Prof. G.S. Ladde  
Professor  
The University of Texas at Arlington  
Department of Mathematics  
(817)273-3261  
e-mail:

Prof. A. Larhtaria  
Associate Professor  
Pennsylvania State University  
Department of Engineering Science & Mechanics  
(814)863-4319  
e-mail:

Mrs. Bonita A. Lawrence  
University of Texas at Arlington  
Department of Mathematics  
(817)273-3261  
e-mail:

Prof. Dening Li  
Associate Professor  
West Virginia University  
Mathematics Department  
(304)293-2011 x308  
e-mail: li@math.wvu.edu

Prof. Wing Kam Liu  
Professor  
Northwestern University  
Mechanical Engineering Department  
(708)491-7094  
e-mail: wkl@andre.mech.nwu.edu

Dr. Ling Ma  
Professor  
Carnegie Mellon University  
Department of Mathematics  
(412)268-2545  
e-mail: maling@andrew.cmu.edu

Dr. Andrei Malevsky

**University of Minnesota  
Army High Performance Computing Research Center  
(612)626-8066  
e-mail: andy@klissy@msi.umn.edu**

**Prof. Joseph D. Myers  
Associate Professor  
U.S. Military Academy  
Department of Math Sciences  
(914)938-5611  
e-mail: aj5831@usma2.usma.edu**

**Dr. Julian Palmore  
Professor  
University of Illinois  
Department of Mathematics  
(217)352-6511 x681  
e-mail: palmore@osiris.cso.uiuc.edu**

**Major Anne Patenaude  
Military Assistant to Deputy Under Secretary of the Army  
(703)697-0366  
e-mail:**

**Dr. Felipe Pereira  
Research Assistant Professor  
Purdue University  
Department of Mathematics  
(317)494-1921  
e-mail: pereira@math.purdue.edu**

**Mr. John Petty  
United States Military Academy  
Department of Mathematical Sciences  
(914)938-3073  
e-mail: aa2095@usma2.usma.edu**

**Dr. Louis Piscitelle  
Research Mechanical Engineer  
U.S. Army Natick RD & E Center  
(508)651-5078  
e-mail:**

**Prof. M. Potasek  
Professor  
Columbia University  
Department of Applied Physics**

**e-mail:**

**Prof. Fernando Reitich**  
**Professor**  
**Carnegie Mellon University**  
**(412)268-2545**  
**e-mail: reitich@andrew.cmu.edu**

**Dr. Rouben Rostamian**  
**National Science Foundation**  
**(202)357-3686**  
**e-mail:**

**Mr. Keith Saints**  
**Student**  
**Cornell University**  
**Center for Applied Mathematics**  
**(607)255-3399**  
**e-mail: keith@macomb.tn.cornell.edu**

**Dr. S. Sathananthan**  
**Assistant Professor of Mathematics**  
**Tennessee State University**  
**Department of Physics, Mathematics and Computer Science**  
**(615)251-1225**  
**e-mail:**

**Prof. George Sell**  
**Professor/Director**  
**University of Minnesota**  
**AHPCRC**  
**(612)626-8080**  
**e-mail:**

**Mr. B. Shanker**  
**Pennsylvania State University**  
**Department of Engineering Science & Mechanics**  
**(814)863-0998**  
**e-mail:**

**Dr. Naveen Sharma**  
**Research Associate**  
**Kent State University**  
**Department of Mathematics & Computer Science**  
**(216)672-4004 x252**  
**e-mail: sharma@mcs.kent.edu**

**Prof. Michael Shearer**  
Professor  
North Carolina State University  
Department of Mathematics  
(919)515-3298  
e-mail: [shearer@math.ncsu.edu](mailto:shearer@math.ncsu.edu)

**Dr. L.S. Shieh**  
Professor  
University of Houston  
Department of Electrical Engineering  
(713)743-4439  
e-mail: [ELEEF6@Jetson.UH.edu](mailto:ELEEF6@Jetson.UH.edu)

**Mr. Kiran Shivaswamy**  
Master's Candidate  
Colorado State University  
Mechanical Engineering Department  
(303)491-7479  
e-mail: [kira@carbon.lance.colostate.edu](mailto:kira@carbon.lance.colostate.edu)

**Prof. Chi-Wang Shu**  
Associate Professor  
Brown University  
Department of Applied Mathematics  
(401)863-2449  
e-mail: [shu@cfm.brown.edu](mailto:shu@cfm.brown.edu)

**Mr. Royce Soanes**  
Mathematician  
Watervliet Arsenal  
( )266-5907  
e-mail:

**Dr. Eduardo Socolovsky**  
Associate Professor  
Hampton University  
Center for Nonlinear Analysis, Mathematics  
(804)727-5030  
e-mail: [eduardo@gamma.math.hampton.edu](mailto:eduardo@gamma.math.hampton.edu)

**Prof. H. Mete Soner**  
Professor  
Carnegie Mellon University  
(412)268-2545  
e-mail: [hs0w@andrew.cmu.edu](mailto:hs0w@andrew.cmu.edu)

**Dr. Janet Spoonamore**  
**Acting Chief**  
**USA-CERL**  
**Facility Management Division**  
**(217)373-7267**  
**spoonam@card.uiuc.edu**

**Mr. Anuj Srivastava**  
**Graduate Student**  
**Washington University**  
**Department of Electrical Engineering**  
**(314)935-5569**  
**e-mail: anuj@hyperion.wustl.edu**

**Mr. Kjartan Stefansson**  
**Cornell University**  
**Department of Computer Science**  
**(607) 255-1187**  
**e-mail: stefan@cs.cornell.edu**

**Prof. Moss Sweedler**  
**Cornell University**  
**Math Science Institute**  
**(607)255-4373**  
**e-mail: Moss\_Sweedler@cornell.edu**

**Prof. John Tsitsiklis**  
**Professor**  
**MIT**  
**Department of Electrical Engineering & Computer Science**  
**(617)253-6175**  
**e-mail: jnt@athena.mit.edu**

**Ms. Judith Underwood**  
**Graduate Student**  
**Cornell University**  
**Department of Computer Science**  
**(607)255-4934**  
**e-mail: under@cs.cornell.edu**

**Dr. John D. Vasilakis**  
**Chief**  
**Benet Laboratories**  
**Research Division**  
**(518)266-5904**  
**e-mail: basilaki@pica.army.mil**



**Dr. Stephanos Venakides**  
Professor  
Duke University  
Department of Mathematics  
(919)660-2815  
e-mail: ven@math.duke.edu

**Prof. Mladen A. Vouk**  
Associate Professor  
North Carolina State University  
Department of Computer Science  
(919)515-7886  
e-mail: vouk@adm.csc.ncsu.edu

**Dr. Paul S. Wang**  
Director of Research  
Kent State University  
Department of Mathematics & Computer Science  
(216)672-4004 x110  
e-mail: pwang@mcs.kent.edu

**Dr. Richard A. Weiss**  
Physicist  
U.S. Army Engineer Waterways Experiment Station  
(601)634-2194  
e-mail:

**Prof. R. E. White**  
Professor  
North Carolina State University  
Department of Mathematics  
(919)515-7678  
e-mail: white@fx40.math.ncsu.edu

**Prof. John R. Whiteman**  
Vice Principal  
Brunel University  
Institute of Computational Mathematics  
44-895-203270  
e-mail: John.Whiteman@brunel.ac.uk

**Dr. Julian J. Wu**  
Professor  
Army Research Office  
Mathematical & Computer Sciences Division  
(919)549-4332  
e-mail: JJWU@brl.mil

**Mr. Scott A. Wymer**  
**Pennsylvania State University**  
**Department of Engineering Science & Mechanics**  
**(814)863-0998**  
**e-mail:**

**Prof. Stephen S.-T. Yau**  
**Director of Control & Information Laboratory**  
**University of Illinois at Chicago**  
**Department of Mathematics, Statistics & Computer Science**  
**(312)996-3065**  
**e-mail: U32790@UICVM.BITNET**

**Prof. David Yuen**  
**Professor**  
**University of Minnesota**  
**Minnesota Supercomputer Institute**  
**(612)624-1868**  
**e-mail: davey@sop.geo.umn.edu**

**Prof. Qiang Zhang**  
**Assistant Professor**  
**SUNY at Stony Brook**  
**Department of Applied Mathematics & Statistics**  
**(516)632-7567**  
**e-mail: zhang@ams.sunysb.edu**

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release: Distribution unlimited		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S)  ARO Report 94-1			5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION  Army Research Office		6b. OFFICE SYMBOL (If applicable)  AMXRO-MCS	7a. NAME OF MONITORING ORGANIZATION		
6c. ADDRESS (City, State, and ZIP Code)  P.O. Box 12211 Research Triangle Park, NC 27709-2211			7b. ADDRESS (City, State, and ZIP Code)		
8a. NAME OF FUNDING / SPONSORING ORGANIZATION  AMSC		8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER		
8c. ADDRESS (City, State, and ZIP Code)			10. SOURCE OF FUNDING NUMBERS		
			PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.
					WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification)  Transactions of the Eleventh Army Conference on Applied Mathematics and Computing.					
12. PERSONAL AUTHOR(S)					
13a. TYPE OF REPORT Technical Report		13b. TIME COVERED FROM 01-93 TO 02-94		14. DATE OF REPORT (Year, Month, Day) 1994 March	
15. PAGE COUNT 731					
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	Fluid and solid mechanics, mathematical physics and numerical methods, symbolic computation, control theory, and Stochastic techniques.		
19. ABSTRACT (Continue on reverse if necessary and identify by block number)  (U) This is a technical report resulting from the Eleventh Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat many Army applied mathematical problems.					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS				21. ABSTRACT SECURITY CLASSIFICATION	
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Francis G. Dressel			22b. TELEPHONE (Include Area Code) (919) 549-4319		22c. OFFICE SYMBOL AMXRO-MCS